DEVELOPING NOVEL CRITERIA TO CLASSIFY ARDS SEVERITY USING A MACHINE LEARNING APPROACH

Mohammed Gamal Sayed Abdelall

UNIVERSITAT ROVIRA i VIRGILI

Developing Novel Criteria to Classify ARDS Severity using a Machine Learning Approach

MOHAMMED GAMAL SAYED ABDELALL

DOCTORAL THESIS
2022

Developing Novel Criteria to Classify ARDS Severity using a Machine Learning Approach

DOCTORAL THESIS

Author:

MOHAMMED GAMAL SAYED ABDELALL

Advisor:

Dr. David Riaño Ramos

Departament d'Enginyeria Informàtica i Matemàtiques (DEIM)

UNIVERSITAT ROVIRA i VIRGILI

Tarragona

2022

UNIVERSITAT ROVIRA i VIRGILI

Departament d'Enginyeria Inform`atica i Matem`atiques (DEIM)
Av. Paisos Catalans, 26
43007 Tarragona, Spain

I STATE that the present study, entitled "Developing Novel Criteria to Classify ARDS Severity using a Machine Learning Approach", presented by MOHAMMED GAMAL SAYED ABDELALL, for the award of the degree of Doctor, has been carried out under my supervision at the Departament d'Enginyeria *Informàtica i Matemàtiques* (DEIM), *Universitat Rovira i Virgili* (URV).

Tarragona, Spain, May 2022.

Doctoral Thesis Supervisor,

Dr. David Riaño Ramos

Dedicated To The SPIRIT Of My FATHER

Abstract

Acute respiratory distress syndrome (ARDS) is a noncardiogenic pulmonary edema, lung inflammation with hypoxemia, and decreased lung compliance. ARDS is a heterogeneous syndrome with a fatal outcome, a constellation of clinical and physiologic observations thought to represent a common pathology. Pathogenesis of ARDS remains elusive, and there is no gold standard diagnostic test. There is a lot of heterogeneity in ARDS diagnosis, the possibility that ARDS is, in fact, a collection of different diseases that have not yet been separately identified. In addition, the disease trajectory of patients within each ARDS category can impact outcome. Most ARDS patients require mechanical ventilation (MV).

In front of the medical difficulties to properly address ARDS issues, as they are reported in multiple specialized publications, in this thesis we hypothesized that the use of modern machine learning (ML) technologies could improve our knowledge and our capacity to predict and address these ARDS issues. In order to achieve these objectives (i) we proposed a novel formula [$PaO_2/(FiO_2xPEEP)$ or P/FP_E] for PEEP≥5 and corresponding cut-off values to address Berlin's definition gap for ARDS severity by using ML approaches. We examined P/FP_E values delimiting the boundaries of mild, moderate, and severe ARDS. We applied ML to predict ARDS severity after onset over time by comparing current Berlin PaO_2/FiO_2 criteria with P/FP_E under three different scenarios, (ii) we aimed at characterizing the best early scenario during the first two days in the intensive care unit (ICU) to predict MV duration after ARDS onset using ML approaches, and (iii) we validated P/FP_E as a predictor of ICU mortality beyond the current state of the art using intuitive classification thresholds based on ML.

We extracted clinical data from the first 3 ICU days after ARDS onset from the single-center MIMIC-III critical care database (MetaVision, 2008-2012) and the multicenter

eICU Collaborative Research Database across the United States between 2014 and 2015. Disease progression in each database was tracked along the first 3 ICU days to assess ARDS severity. We included variables of arterial oxygenation and ventilator settings that were readily available in routine clinical practice to guarantee clinical relevance within a wide range of ARDS severity. Three robust ML techniques were implemented using Python 3.7: LightGBM, RF, and XGBoost.

We proved that our novel P/FP_E index to assess ARDS severity after onset over time is markedly better than current PaO_2/FiO_2 ratio. The best early MV duration prediction model was obtained with data captured in the 2^{nd} day. ML models might have important implications for optimizing ICU resource utilization and high acute cost reduction of MV. Moreover, P/FP_E index is a more sensitive predictor of ICU mortality over time than the PaO_2/FiO_2 ratio in all ARDS categories.

ARDS is a predominantly clinical diagnosis, but there have been difficulties in agreeing on a standardized, universal definition. The PaO_2/FiO_2 ratio classifies the severity of ARDS based on the degree of the oxygenation deficit and within each of those categories, patients are assumed to be less heterogeneous. However, recent observations showed that mild ARDS is underappreciated because it had a high mortality rate. Since the current Berlin definition for ARDS does not account for PEEP in its calculation, it provides an incomplete picture of actual ARDS severity. This thesis provides a solution for that dilemma by using the P/FP_E index, taking applied PEEP into account and creating three grades of severity based on intuitive classification thresholds that are different from the Berlin definition, and within each of these ARDS categories, the patients had a similar degree of lung severity. We believe this thesis is an important addition to the current ARDS story.

***Keywords**:* Machine Learning, Artificial Intelligence, Decision Support Systems, Computer-Based Clinical Prediction Modelling, Performance Assessment, ARDS Severity, Intensive Care Unit, Mortality, Mechanical Ventilation

Resumen (Spanish)

El síndrome de dificultad respiratoria aguda (ARDS) es un edema pulmonar no cardiogénico, inflamación pulmonar con hipoxemia y distensibilidad pulmonar disminuida. El ARDS es un síndrome heterogéneo con un desenlace fatal, una constelación de observaciones clínicas y fisiológicas que se cree que representan una patología común. La patogenia del ARDS sigue siendo esquiva y no existe una prueba diagnóstica estándar de oro. Hay mucha heterogeneidad en el diagnóstico de ARDS, la posibilidad de que ARDS sea, de hecho, una colección de diferentes enfermedades que aún no se han identificado por separado. Además, la trayectoria de la enfermedad de los pacientes dentro de cada categoría de ARDS puede afectar el resultado. La mayoría de los pacientes con ARDS requieren ventilación mecánica (MV).

Frente a las dificultades médicas para abordar adecuadamente los problemas de ARDS, tal como se reportan en múltiples publicaciones especializadas, en esta tesis planteamos la hipótesis de que el uso de tecnologías modernas de aprendizaje automático (ML) podría mejorar nuestro conocimiento y nuestra capacidad para predecir y abordar estos ARDS. cuestiones. Para lograr estos objetivos (i), propusimos una fórmula novedosa [$PaO_2/(FiO_2xPEEP)$ or P/FP_E] para PEEP≥5 y los valores de corte correspondientes para abordar la brecha de definición de Berlín para la gravedad del ARDS mediante el uso de enfoques ML. Examinamos los valores de P/FP_E que delimitan los límites del ARDS leve, moderado y grave. Aplicamos ML para predecir la gravedad del ARDS después del inicio a lo largo del tiempo comparando los criterios actuales de PaO_2/FiO_2 de Berlín con P/FP_E en tres escenarios diferentes, (ii) apuntamos a caracterizar el mejor escenario temprano durante los dos primeros días en la unidad de cuidados intensivos (ICU) para predecir la duración de la MV después del inicio del ARDS utilizando enfoques de ML, y (iii) validamos P/FP_E como predictor de mortalidad en la ICU más allá del estado actual del arte utilizando umbrales de clasificación intuitivos basados en ML.

Extrajimos datos clínicos de los primeros 3 días en la UCI después del inicio del ARDS de la base de datos MIMIC-III de cuidados críticos de un solo centro (MetaVision, 2008-2012) y la base de datos de investigación colaborativa eICU multicéntrica en los Estados Unidos entre 2014 y 2015. Se realizó un seguimiento de la progresión de la enfermedad en cada base de datos a lo largo de esos 3 días en la ICU para evaluar la gravedad del ARDS. Incluimos variables de oxigenación arterial y configuración del ventilador que estaban fácilmente disponibles en la práctica clínica habitual para garantizar la relevancia clínica dentro de un amplio rango de gravedad del ARDS. Se implementaron tres técnicas robustas de ML utilizando Python 3.7: LightGBM, RF y XGBoost.

El nuevo índice P/FP_E para evaluar la gravedad del ARDS después del inicio con el tiempo es notablemente mejor que la relación PaO_2/FiO_2 actual. El mejor modelo de predicción de duración temprana de MV se obtuvo con datos capturados en el segundo día. Los modelos de ML pueden tener implicaciones importantes para optimizar la utilización de los recursos de la ICU y la reducción de los altos costos agudos de la MV. El índice P/FP_E es un predictor más sensible de la mortalidad en la ICU a lo largo del tiempo que la relación PaO_2/FiO_2 en todas las categorías de ARDS.

El ARDS es un diagnóstico predominantemente clínico, pero ha habido dificultades para acordar una definición universal estandarizada. El cociente PaO_2/FiO_2 clasifica la gravedad del ARDS en función del grado de déficit de oxigenación y, dentro de cada una de esas categorías, se supone que los pacientes son menos heterogéneos. Sin embargo, observaciones recientes mostraron que el ARDS leve se subestima porque tiene una alta tasa de mortalidad. Dado que la definición actual de Berlín para ARDS no tiene en cuenta la PEEP en su cálculo, proporciona una imagen incompleta de la gravedad real de ARDS. Nuestra tesis proporciona una solución para ese dilema utilizando el índice P/FP_E, teniendo en cuenta la PEEP aplicada y creando tres grados de gravedad basados en umbrales de clasificación intuitivos que son diferentes a la

definición de Berlín, y dentro de cada una de estas categorías de ARDS, los pacientes tenían un grado similar de gravedad pulmonar. Creemos que nuestra tesis es una adición importante a la historia actual de ARDS.

Palabras clave: Aprendizaje automático, inteligencia artificial, sistemas de soporte a la decisión, modelado de predicción clínica basado en computadora, evaluación del desempeño, gravedad del ARDS, unidad de cuidados intensivos, mortalidad, ventilación mecánica

Resum (Catalan)

La síndrome d'angoixa respiratòria aguda (ARDS) és un edema pulmonar no cardiogènic, inflamació pulmonar amb hipoxèmia i disminució del compliment pulmonar. El ARDS és una síndrome heterogènia amb un desenllaç fatal, una constel·lació d'observacions clíniques i fisiològiques que es creu que representen una patologia comuna. La patogènesi de l'ARDS continua essent esquiva i no hi ha cap prova diagnòstica estàndard. Hi ha molta heterogeneïtat en el diagnòstic del ARDS, la possibilitat que el ARDS sigui, de fet, un conjunt de diferents malalties que encara no s'han identificat per separat. A més, la trajectòria de la malaltia dels pacients dins de cada categoria de ARDS pot afectar el resultat. La majoria dels pacients amb ARDS requereixen ventilació mecànica (MV).

Davant de les dificultats mèdiques per abordar correctament els problemes del ARDS, tal com es reporten en múltiples publicacions especialitzades, en aquesta tesi hem plantejat la hipòtesi que l'ús de tecnologies modernes d'aprenentatge automàtic (ML) podria millorar el nostre coneixement i la nostra capacitat per predir i abordar aquests ARDS. qüestions. Per assolir aquests objectius (i) vam proposar una fórmula nova [$PaO_2/(FiO_2xPEEP)$ or P/FP_E] per a PEEP≥5 i els valors de tall corresponents per abordar la bretxa de definició de Berlín per a la gravetat de l'ARDS mitjançant enfocaments ML. Es van examinar els valors de P/FP_E que delimiten els límits de l'ARDS lleu, moderat i greu. Hem aplicat ML per predir la gravetat del ARDS després de l'aparició al llarg del temps comparant els criteris actuals de PaO_2/FiO_2 de Berlín amb P/FP_E en tres escenaris diferents, (ii) vam tenir com a objectiu caracteritzar el millor escenari precoç durant els dos primers dies a la unitat de cures intensives (ICU) per predir la durada de la MV després de l'inici de l'ARDS mitjançant enfocaments de ML, i (iii) vam validar P/FP_E com a predictor de la mortalitat de la ICU més enllà de l'estat actual de la tècnica mitjançant llindars de classificació intuïtius basats en ML.

Vam extreure dades clíniques dels primers 3 dies de l'UCI després de l'inici de l'ARDS de la base de dades MIMIC-III de cures crítiques d'un sol centre (MetaVision, 2008-2012) i la base de dades de recerca col·laborativa multicèntrica d'eICU als Estats Units entre 2014 i 2015. Es va fer un seguiment de la progressió de la malaltia a cada base de dades al llarg d'aquests 3 dies d'ICU per avaluar la gravetat de l'ARDS. Es van incloure variables d'oxigenació arterial i configuració del ventilador que estaven fàcilment disponibles a la pràctica clínica rutinària per garantir la rellevància clínica dins d'un ampli rang de gravetat del ARDS. Es van implementar tres tècniques de ML robustes mitjançant Python 3.7: LightGBM, RF i XGBoost.

El nou índex P/FP_E per avaluar la gravetat de l'ARDS després de l'aparició al llarg del temps és notablement millor que la relació PaO_2/FiO_2 actual. El millor model de predicció de la durada de la MV primerenca es va obtenir amb dades capturades el segon dia. Els models de ML poden tenir implicacions importants per optimitzar la utilització dels recursos de la ICU i la reducció de costos aguda de la MV. L'índex P/FP_E és un predictor més sensible de la mortalitat a l'ICU al llarg del temps que la relació PaO_2/FiO_2 en totes les categories d'ARDS.

El ARDS és un diagnòstic predominantment clínic, però hi ha hagut dificultats per posar-se d'acord en una definició estandarditzada i universal. La relació PaO_2/FiO_2 classifica la gravetat de l'ARDS en funció del grau de dèficit d'oxigenació i dins de cadascuna d'aquestes categories, se suposa que els pacients són menys heterogenis. No obstant això, observacions recents van demostrar que el ARDS lleu no es valora perquè tenia una alta taxa de mortalitat. Com que la definició actual de Berlín per a l'ARDS no té en compte la PEEP en el seu càlcul, proporciona una imatge incompleta de la gravetat real de l'ARDS. La nostra tesi proporciona una solució a aquest dilema mitjançant l'ús de l'índex P/FP_E, tenint en compte la PEEP aplicada i creant tres graus de gravetat basats en llindars de classificació intuïtius que són diferents de la definició de Berlín, i dins de cadascuna d'aquestes categories de ARDS, els pacients tenien

un grau similar de gravetat pulmonar. Creiem que la nostra tesi és una addició important a la història actual de l'ARDS.

Paraules clau: Aprenentatge automàtic, intel·ligència artificial, sistemes de suport a la presa de decisions, modelització de predicció clínica basada en ordinador, avaluació del rendiment, gravetat del ARDS, unitat de cures intensives, mortalitat, ventilació mecànica

Acknowledgments

I would like to express my gratitude to my supervisor Dr. David Riaño Ramos for his useful guidance, insightful comments, and considerable encouragement to complete this thesis. He has guided me to pursue important problems that will have a practical impact and were always available to guide me whenever I approached him. This work would not have been completed without his encouragement and patience.

Also, I would like to express my gratitude to Romain Pirracchio, Jesús Villar, Arthur S. Slutsky, Yub Raj Sedhai, Abdul Hameed Asif M., Mir Wasey Ali Yadullahi, Shrestha Dhan, Acharya Roshan, and Piyush Mathur for their support and assistance. They provided critical appraisal during data analysis and interpretation. Without their precious support it would not be possible to conduct this research.

Furthermore, I would like to acknowledge all the funding institutions that have supported this research: including the Martí-Franquès Research Fellowship Programme - Doctoral Grant provided by Universitat Rovira i Virgili.

Finally, I must express my very profound gratitude to my family for providing me with unfailing support and continuous encouragement throughout my years of study and throughout the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Author

MOHAMMED GAMAL SAYED ABDELALL

TABLE OF CONTENTS

Nomenclature

AECC	American–European Consensus Criteria
AI	Artificial Intelligence
ARDS	Acute Respiratory Distress Syndrome
AUC	Area Under the Curve
CI	Confidence Interval
CORR	Correlation between the predicted and actual values of the target variable
ECMO	Extracorporeal Membrane Oxygenation
eICU	eICU Collaborative Research Database
FiO_2	Fraction of the Oxygen in the inspired air
FPR	False Positive Rate
HR	Heart Rate
GOSS	Gradient-based One-Side Sampling
LightGBM	Light Gradient Boosting Machine
ICU	Intensive Care Unit
ML	Machine Learning
MIMIC-III	Medical Information Mart for Intensive Care Database
MV	Mechanical Ventilation
$PaCO_2$	Partial Pressure of Arterial CO_2
PaO_2	Arterial Oxygen Tension

PaO_2/FiO_2	Ratio of Partial Pressure of Arterial O_2 to Fraction of inspired O_2
P/FP_E	New ARDS Severity Criteria
PEEP or P_E	Positive End-Expiratory Pressure
RF	Random Forest
RMSE	Root Mean Square Error
ROC	Receiver Operating Characteristics
ROSE	Reevaluation of Systemic Early Neuromuscular Blockade
RR	Respiratory Rate
SD	Standard Deviation
SpO_2	Oxygen Saturation
TPR	True Positive Rate
XGBoost	eXtreme Gradient Boosting

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Introduction

Acute respiratory distress syndrome (ARDS) is an acute and intense inflammatory disease process of the lungs with an associated high mortality rate of about 40% in non-COVID-19 ARDS patients [1,2]. ARDS is a highly heterogeneous syndrome without a specific diagnostic test [3-5]. According to the LUNG-SAFE study, ARDS is unrecognized in more than half of patients at the time of fulfillment of ARDS criteria [1]. The current “Berlin definition” of ARDS and severity grades is under controversy [5-8], mainly because it does not assess the "*true*" severity of lung injury, which hinders its clinical utility. The previous definition of the American-European Consensus Conference (AECC, 1994) [9] and the Berlin definition are predominantly based on the value of the PaO_2/FiO_2 ratio at the time of ARDS onset [10].

A working definition of ARDS is essentially required for clinical trials, epidemiologic studies, and biological studies. Moreover, a definition of ARDS is required for clinicians to initiate treatments that would improve clinical outcomes [11], although stratification of ARDS -as defined by the Berlin criteria- has been shown to be not very useful for assessing lung severity [8,12]. The empirical PaO_2/FiO_2 cut-offs for “severity” of 100, 200, and 300 mmHg are arbitrary and poorly validated [13]. A recently published Reevaluation of Systemic Early Neuromuscular Blockade (ROSE) trial emphasized the variability of these PaO_2/FiO_2 cut-offs as the investigators did not enroll patients based on the PaO_2/FiO_2 at the time of ARDS onset, but based on a PaO_2/FiO_2<150 mmHg within the first 48-h after ARDS diagnosis [14,15].

The PaO_2/FiO_2 ratio strongly depends on ventilator settings, including positive end-

expiratory pressure (PEEP), inspiratory:expiratory time (I:E) ratio, and FiO_2, and the requirement of a minimum PEEP of 5 cmH_2O did not substantially improve Berlin prediction compared to AECC [13,16]. Besides, Berlin definition does not account for the nonlinear relationship of PaO_2 and FiO_2 [17] and has a limited predictive accuracy in recent trials [18-21].

Assessment of severity in ARDS still faces many challenges. **In an attempt to handle them**, this thesis proposes a novel criterion for ARDS severity and develops several machine learning (ML) approaches for the "*true*" ARDS severity classification, and the prediction of the duration of mechanical ventilation (MV) and mortality of ARDS patients with models that outperform the current state of the art.

1.2. Objectives and scope of the research

The main objectives of this thesis are:

1. To propose a novel criterion to classify ARDS severity using ML approach.
2. To propose a ML-based model for predicting duration of MV in ARDS.
3. To assess the efficacy of the proposed ARDS criteria with ML techniques for mortality prediction.

1.3. Main contributions of the thesis

The main contributions of this thesis are the following:

- We proposed a new P/FP_E criterion for ARDS severity and examined P/FP_E values delimiting the boundaries of mild, moderate, and sever ARDS. We applied ML to predict ARDS severity after onset over time by comparing current PaO_2/FiO_2 ratio with P/FP_E index under three different scenarios.

The results of the previous studies have been published in the following "Q1" journal:

***Sayed M**, Riaño D, Villar J. Novel criteria to classify ARDS severity using a machine learning approach. Crit Care. 2021;25(1):150. doi: 10.1186/s13054-021-03566-w. PMID: 33879214; PMCID: PMC8056190.*

- We applied ML to develop novel models to predict the duration of MV in ARDS after onset over time under three different clinical scenarios.

The results of the previous studies have been published in the following "Q1" journal:

***Sayed M**, Riaño D, Villar J. Predicting Duration of Mechanical Ventilation in Acute Respiratory Distress Syndrome Using Supervised Machine Learning. J Clin Med. 2021;10(17):3824. doi: 10.3390/jcm10173824. PMID: 34501270; PMCID: PMC8432117.*

- We analyze of the effectiveness of the proposed ARDS criteria with ML techniques for mortality prediction. Specifically, we use a random forest (RF) algorithm to develop models predicting intensive care unit (ICU) mortality for each severity grade of ARDS using routinely values of gas-exchange and ventilator settings.

The results of the previous studies have been submitted to the following "Q1" journal:

***Sayed M**, Riaño D, Abdul Hameed Asif M, Mir Wasey Ali Yadullahi, Shrestha Dhan, Acharya Roshan, Sedhai Yub Raj. Validating P/FP_E index to predict mortality in acute respiratory distress syndrome using machine learning. Ann Am Thorac Soc. 2022 April. Submitted with Manuscript ID: White-202204-359OC.*

- **Other publications derived from the thesis:**

***Sayed, M.** & Riaño, D. (2019). Modelling ICU Patients to Improve Care Requirements and Outcome Prediction of Acute Respiratory Distress Syndrome: A Supervised Learning Approach. In Artificial Intelligence in Medicine: Knowledge Representation and Transparent and Explainable Systems (pp. 39-49). Springer, Cham. https://doi.org/10.1007/978-3-030-37446-4_4*

1.4. Thesis organization

The thesis is structured into the following chapters:

- **Chapter 1** includes the motivation behind the thesis, the structure of the thesis, and the contributions it makes.
- **Chapter 2** describes the background to ARDS assessment, ML techniques, statistical analysis, datasets, and evaluation metrics.
- **Chapter 3** presents the proposed ARDS criteria for ARDS severity classification.
- **Chapter 4** presents MV duration prediction approach using ML.
- **Chapter 5** provides the analysis of the efficacy of the proposed ARDS criteria with ML techniques for mortality prediction.
- **Chapter** 6 provides the conclusions of the thesis and some lines of future research.

CHAPTER 2

Background

2.1 Acute Respiratory Disease Syndrome (ARDS)

ARDS is an acute and intense inflammatory disease process of the lungs with an associated high mortality rate of about 40% in non-COVID-19 ARDS patients [1,2]. ARDS is a highly heterogeneous syndrome without a specific diagnostic test [3,4]. According to the LUNG-SAFE study, ARDS is unrecognized in more than half of patients at the time of fulfillment of ARDS criteria [1]. ARDS definitions have varied over time. ARDS was first described by Ashbaugh et al. in 1967. Then, for more than twenty years, there was no standard definition of ARDS. The 1994 AECC definition became widely accepted, but had limitations [5]. The current "Berlin definition" is under controversy [5-8]. This is due to that Berlin definition does not solve the problems with the AECC definition. Furthermore, The stratification of ARDS patients as proposed by the Berlin definition is useless for assessing the "*true*" severity of lung injury and hinders enrolling patients into clinical trials [8]. Of note, the previous AECC [9] and the Berlin definitions are predominantly based on the value of the PaO_2/FiO_2 ratio at the time of ARDS onset [10].

A working definition of ARDS is essentially required for clinical trials, epidemiologic studies, and biological studies. Moreover, a practical definition of ARDS is required for clinicians to initiate treatments that would improve clinical outcomes [11], although stratification of ARDS -as defined by the Berlin criteria- has been shown not very useful for assessing lung severity [8,12]. The empirical PaO_2/FiO_2 cut-offs for "severity" of 100, 200, and 300 mmHg are arbitrary and poorly validated [13]. The Berlin definition of ARDS

identifies three mutually exclusive categories of lung severity with PaO_2/FiO_2 ratios in the ranges >200-300 mmHg (mild ARDS), >100-200 (moderate ARDS), and ≤100 (severe ARDS) [6,7]. Some studies [8,9] have reported a progression of costs from mild to moderate to severe ARDS. Despite global acceptance of Berlin criteria [10], some specialists have questioned about its ability to assess the "*true"* severity of lung injury [11].

A recently published ROSE trial emphasized the variability of these PaO_2/FiO_2 cut-offs as the investigators did not enroll patients based on the PaO_2/FiO_2 at the time of ARDS onset, but based on a PaO_2/FiO_2<150 mmHg within the first 48-h after ARDS diagnosis [14,15]. The PaO_2/FiO_2 ratio strongly depends on ventilator settings, including positive end-expiratory pressure (PEEP), inspiratory:expiratory time (I:E) ratio, and FiO_2, and the requirement of a minimum PEEP of 5 cmH_2O did not substantially improve Berlin prediction compared to AECC [13,16]. Besides, Berlin definition does not account for the nonlinear relationship of PaO_2 and FiO_2 [17] and has a limited predictive accuracy in recent trials [18-21].

ARDS is an important cause of morbidity, mortality, and costs in ICUs worldwide [22]. It is a life-threatening form of acute respiratory failure characterized by inflammatory pulmonary edema leading to severe hypoxemia, requiring endotracheal intubation and MV in most cases [23]. The number of days on MV during the ICU stay is a major driver of high acute care costs [24-26]. There is a believe that an important intervention to mitigate these costs is timely recognition and treatment of conditions that can cause serious complications.

2.2 Machine Learning Techniques

ML is a subset of artificial intelligence (AI) in which machines extract knowledge from the data provided. ML is an exploratory process where there is no one-method-fits-all solution. ML merges statistical analysis techniques with computer science to produce algorithms capable

of “statistical learning”. ML algorithms are divided into two categories: supervised and unsupervised. Supervised learning algorithms, as the ones used in the thesis, detect relationship between potential explanatory features and a known target outcome. ML and data-driven approaches are gaining traction in various application fields, as for example healthcare decision support. Research on decision support applications in healthcare, such as those related to diagnosis, prediction, etc., has seen a surge interest over recent years. This development is due to increasing in the availability of data as well as to advances in ML research. As the name says, these techniques rely on data availability to extract knowledge and train algorithms. Decision support in healthcare based on ML is a promising field. For over fifty years, the possibilities of ML to support healthcare professionals in decision making have been investigated, and several prototypes and real products have been deployed in real-world. There are numerous decisions that a ML-based algorithm might assist with. For instance, the diagnosis of a disease severity in ICU. This is a classification task (disease severity A vs. disease severity B). In this case, the users are critical care physicians. Other decision-making tasks can include predicting ICU resource needs (e.g., duration of MV) and predicting ICU outcome (e.g., mortality) [27,28].

In this thesis, we employed three out-of-the-shelf -robust ML techniques to construct the prediction models, namely Random Forest (RF) [29], eXtreme Gradient Boosting (XGBoost) [30] and Light Gradient Boosting Machine (LightGBM) [31] are used in this thesis.

RF is a user-friendly, flexible ML technique that, in most circumstances, produces outstanding results without the need to tweak hyperparameters. It is also one of the most extensively utilized algorithms due to its simplicity and versatility. RF can be used for classification and regression problems. RF comprises numerous decision trees. Compared to other traditional classification algorithms, it has low classification error—the number of trees, minimum node size, and number of features used for splitting each node. The main merits of

RF are [29]:

1. It overcomes the problem of overfitting.
2. RF accuracy and variable importance are automatically generated.
3. Produced forests can be saved for future consideration.

Among the ML approaches used in the literature, the gradient tree boosting (i.e., gradient boosting machine) obtained state-of-the-art outcomes in various real-world applications. XGBoost is a scalable, end-to-end tree boosting approach. Gradient boosted decision trees are implemented quickly and efficiently. The XGBoost system is free and open source. The importance of the XGBoost has been extensively acknowledged in a variety of machine learning and data mining challenges. The ability of XGBoost to scale across all conditions is the most important component of its success. It's five times faster than current machine learning methods. Furthermore, it is scalable to billions of samples in distributed or memory-limited environments. The scalability of XGBoost is achieved thanks to the following [30]:

1. XGBoost employs an efficient tree learning approach for handling large and sparse datasets,

2. XGBoost utilizes out-of-core computation to handle millions of samples on a single computer.

LightGBM improves the gradient boosting technique by integrating an autonomous feature selection method and by concentrating on cases with more significant gradients. To filter out data instances and generate a split value, LightGBM employs Gradient-based One-Side Sampling (GOSS). LightGBM is a high-performance algorithm that quickly processes large amounts of data and dispersed data. Microsoft developed it as a free and open-source project. The LightGBM algorithm has several hyperparameters. The hyperparameters have a significant impact on the performance of the LightGBM algorithm. They are often set manually and fine-tuned through trials and errors. The hyperparameters are "num leaves," which is the number of

leaves per tree, "max depth," which is the tree's maximum depth, and "learning rate" [31].

Grid search was used to identify the optimal values of the input parameters for these algorithms. The grid searching approach involves scanning data to determine the best parameters for a machine learning model. Grid-searching is computationally intensive and can take a long time to complete. The grid searching technique will develop a machine learning model on each feasible parameter combination. It goes over each parameter combination, saving a model for each one [27].

2.3 ARDS Model Assessment with ML

Evaluating the predictive performance of ML models involves assessing their discriminatory and calibration accuracy [28]. Figure 2.1 illustrates how to assess machine learning predictive models throughout their life cycle to maximize their utility and use in clinical practice.

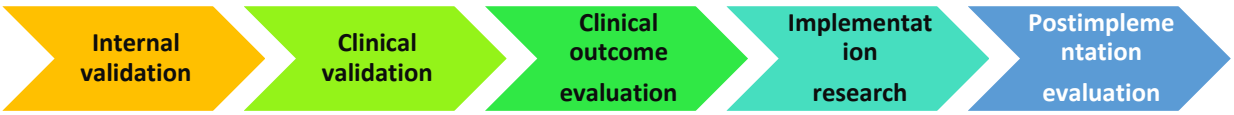


Fig. 2.1. Evaluation life cycle of machine-learned systems in health care [28].

In this thesis, the quality of the prediction models is computed based on a 10-fold cross-validation approach, which means that the dataset was divided into 10 disjoint folds, and in each run, 9 were used for training and the remaining 1 was used for testing. The Area Under the Curve (AUC) and the correlation between the predicted and actual values of severity level (CORR) are used to assess model performance in predicting ARDS severity as a categorical prediction.

In ML, cross-validation is a statistical process for assessing a model's performance on previously unknown data. It is a resampling strategy for testing ML models on a small data

sample to see how they will generally do when used to make predictions on data not included during training [27].

The k-fold cross-validation method includes a single parameter, so-called k, which specifies the number of groups into which the dataset should be split. The following are the basic steps of the k-fold cross-validation method [27]:

1. Randomly shuffle the dataset.
2. Organize the data into k groups.
3. For each distinct group, do:
   3.1 As a holdout or test data set, use the group.
   3.2 As a training data set, use the remaining groupings.
   3.3 Fit a machine learning model to the training set and
       test it against the test set.
   3.4 Keep the evaluation score but toss out the model.
4. Using the sample of model evaluation scores, summarize the model's ability.

The CORR is used in the context of ML to determine the degree of association between two variables. The correlation coefficient is calculated on a scale of [+1,1]. A complete relationship between two variables is represented by a + 1 or a -1. The correlation is positive when one variable rises in lockstep with the other and negative when one falls in lockstep. A CORR value of 0 indicates that there is no association at all. Given two variables x and y, CORR can be expressed as follows [27]:

$$\mathrm{CORR} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{[\sum(x-\bar{x})^2(y-\bar{x})^2]}} \quad (1)$$

where $\bar{x}$ and $\bar{y}$ stand for the mean of x and y, respectively.

The Receiver Operating Characteristics (ROC) curve is an essential evaluation tool for

assessing the effectiveness of any classification algorithm at different threshold levels. As shown in Figure 2.2, the True Positive Rate (TPR) is plotted against the False Positive Rate (FPR), with TPR on the y-axis and FPR on the x-axis to create the ROC curve. The AUC is a scale ranging from 0 to 1 that assesses the separability of classes (for example, patient and healthy). With an AUC of 1, the model is able to distinguish between patients and healthy groups perfectly. The AUC has the privilege of being a commonly reported metric in both medical and recent ML studies. The AUC is a simpler, more generalized metric, to assess the ML performance rather than the varying tradeoffs between sensitivity and specificity. In this metric, 1.00 is considered perfect, and 0.50 is no better than random guessing or flipping a coin; higher values than 0.50 mean that the model is more accurate. However, how high an AUC should be for being "good enough" is application dependent. In some circumstances, an AUC of 0.65 might be very good, while 0.95 might be still rather poor in others. The AUC is a convenient metric as it provides one single number for describing the overall performance and can be used to compare different classifiers. However, it should be kept in mind that it does not say anything about the clinical relevance [27,28].

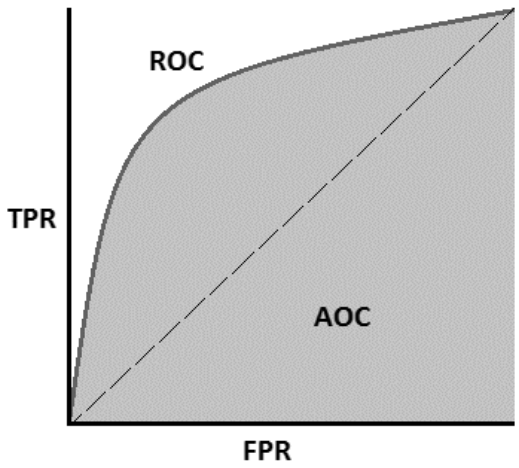


Fig. 2.2. Illustration of the ROC curve.

Besides, the root mean square error (RMSE) is used to evaluate the regression models in this thesis. RMSE can be expressed as follows [27]:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \bar{y}_i)^2}{n}} \quad (2)$$

Where y_i, $\bar{y}_i$ are the actual and the predicted values, and *n* the number of samples, respectively.

2.4 Data sources

In this thesis, we used two publicly available ICU databases: the single-center MIMIC-III dataset [32] and the multicenter eICU dataset [33].

2.4.1 MIMIC-III dataset

Medical Information Mart for Intensive Care database (MIMIC-III) is a publicly available and a large single-center dataset (clinical data of patients admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts) that includes 53,423 distinct ICU admissions for adult patients (age ≥18 years) from 2001 to 2012 [32].

2.4.2 eICU dataset

Telehealth Intensive Care Unit Collaborative Research Database (eICU) is a multicenter ICU dataset with high granular data for more than 200,859 patients' admissions to ICUs monitored by eICU Programs across the United States between 2014 and 2015 [33].

CHAPTER 3

Novel Criteria ARDS Severity Classification

3.1. Introduction

Assessment of severity in ARDS remains a challenge. The relation between oxygenation and prognosis in ARDS varies among published reports [20]. For example, the current mild ARDS category may not be significantly associated with 28-day mortality [34-36]. However, although stratification of severity based on Berlin criteria may be helpful to identify severe ARDS patients, it may have less significance to differentiate between mild and moderate ARDS [20]. A recent study identified two different subgroups of moderate ARDS using a 150 mmHg PaO_2/FiO_2 threshold, and may represent a more homogeneous distribution of ARDS patients across subgroups of severity [37-39]. Whether ARDS outcome relates to severity of respiratory failure [40], a higher severity is a risk factor for prolonged mechanical ventilation [19]. Since PaO_2/FiO_2 does not account for PEEP in its calculation, reported PaO_2/FiO_2 provides a sense of ARDS severity without knowledge of applied PEEP levels.

The main goal of this chapter is to address the first objective of this PhD thesis: Improving the current state of the art by proposing a novel criterion to classify ARDS severity with the use of ML.

Herein, we propose a novel formula [$PaO_2/(FiO_2 x PEEP)$ or P/FP_E] for PEEP≥5 cmH_2O that, together with corresponding thresholds, could serve as an improved criterion to assess ARDS severity beyond current solution. We also aim at determining the thresholds to stratify mild, moderate, and severe ARDS for the new formula. We confirmed that the new P/FP_E

criterion adequately addresses the Berlin's definition gap in computing ARDS severity by including PEEP in the new oxygenation ratio. Increasing the PEEP level with the same FiO_2 yields different PaO_2 and SpO_2 [41]. Thus, including PEEP in calculating the degree of oxygenation severity could be better than the current Berlin definition. We examined this hypothesis by applying machine learning (ML) approaches for predicting ARDS severity over time.

3.2 Methodology

3.2.1 Study Design and Patient Populations

Two critical care databases were used for testing that P/FP_E outperforms the current clinical criteria to stratify the ARDS severity levels, represented by the Berlin definition. Data of the first 3 ICU days (considering day 1 for representative data within the first 24 h after ARDS onset, day 2 for data within 24-48 h after onset, and day 3 for data within 48-72 h after onset) were extracted from a single-center database MIMIC-III (MetaVision, 2008-2012) [32] (N=2738, 1519, and 1341 patients, respectively). The median length of an ICU stay (LOS) of all selected ARDS patients in MIMIC-III was 11.29 days (Q1–Q3: 7.85–17.54). Similarly, data of the first 3 ICU days after ARDS onset were extracted from a multicenter database eICU (2014-2015) [33] (N=5153, 2981, and 2326 patients, respectively). The median length of an ICU LOS of all selected ARDS patients in eICU was 11.72 days (Q1–Q3: 6.92–18.84). All selected patients from both databases fulfilled the Berlin criteria for ARDS, and were stratified into mild, moderate, or severe ARDS [6], and received MV for >48 hours [42,43]. Disease progression of ARDS in each database was tracked along those 3 ICU days to assess lung severity. Patients younger than 18 years were excluded. Clinical data of ARDS patients were

extracted from both databases (MIMIC-III and eICU) using Python 3.7. The selection of clinical variables was based on previous studies [1,19,44-47].

MIMIC-III:

The input variables include baseline demographic information such as age; hemodynamic parameters including mean, maximum and minimum heart rate (HR); ventilator parameters including mean, maximum and minimum respiratory rate (RR), SpO_2, and PEEP. The mean and 95% confidence interval of these predictors on the third ICU day after assessing lung severity are presented in Table 3.1. Tables A1 and A2 in "Appendix A" complement this information with the mean and 95% confidence interval of the variables for the patients at 24h and 48h, respectively. The main target variable was ARDS severity (where 0=mild, 1=moderate, and 2=severe). ICU mortality rates for mild, moderate, and severe ARDS patients in MIMIC-III at 72h (Fig. 3.1), but also at 24h and 48h (Figures A1 and A2 in Appendix A) and their duration of MV were also obtained (see Table A3 in Appendix A).

Table 3.1. Input variables and their descriptive statistics in MIMIC-III at 72-h according to PaO_2/FiO_2.

	Mild	**Moderate**	**Severe**	**All**
A. ARDS Patients	***506 (37.73%)***	***678 (50.56%)***	***157 (11.71%)***	***1,341 (100%)***
*B. Descriptive feature– **means and 95% CI***				
Age	61.77 [60.37, 63.17]	60.61 [59.42, 61.79]	60.24 [57.42, 63.07]	61.01 [60.14, 61.87]
PEEP	7.41 [7.11, 7.71]	9.40 [9.06, 9.75]	11.68 [10.83, 12.52]	8.92 [8.68, 9.16]
Heart Rate_Mean	92 [90, 94]	92 [91, 94]	96 [93, 99]	93 [92, 94]
Respiratory Rate_Mean	21 [20, 21]	21 [21, 22]	22 [21, 23]	21 [21, 22]
Heart Rate_Max	114 [112, 116]	114 [112, 116]	120 [116, 124]	115 [113, 116]
Heart Rate_Min	75 [74, 77]	76 [75, 78]	78 [75, 81]	76 [75, 77]
Respiratory Rate_Max	30 [29, 31]	30 [29, 31]	32 [31, 34]	30 [30, 31]
Respiratory Rate_Min	13 [13, 14]	13 [13, 14]	13 [13, 14]	13 [13, 14]
SpO_2_Mean	97 [97, 98]	96 [96, 97]	96 [95, 96]	97 [96, 97]
SpO_2_Max	100 [100, 101]	100 [99, 100]	100 [99, 100]	100 [99, 100]
SpO_2_Min	90 [89, 90]	88 [87, 89]	85 [83, 87]	88 [88, 89]

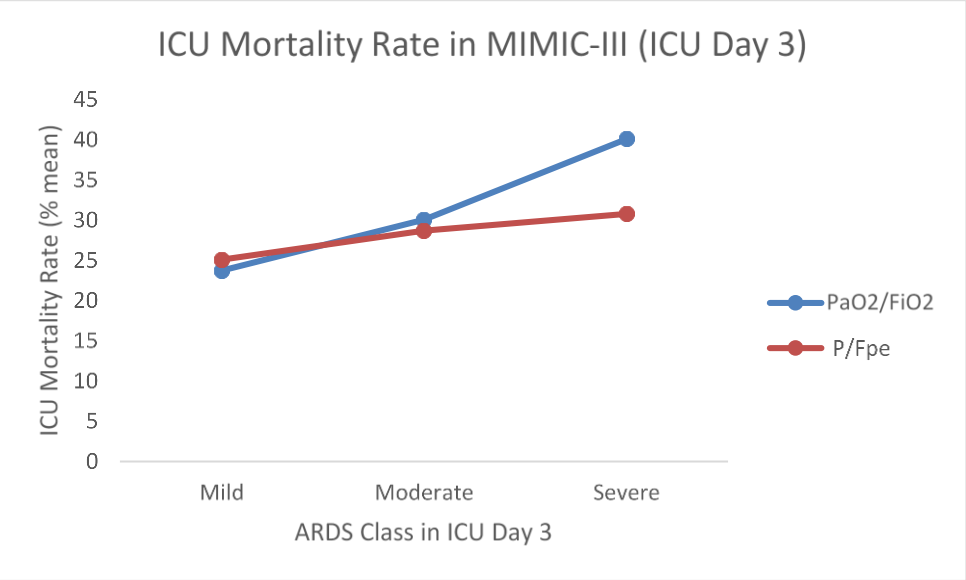


Fig. 3.1. Intensive Care Unit mortality rate at 72 h in relation to degree of lung severity in patients with acute respiratory distress syndrome included in the MIMIC-III database, according to PaO_2/FiO_2 ratio and P/FP_E.

eICU:

Input variables in eICU included baseline demographic information such as age; ventilator parameters including PEEP; blood gas parameters including FiO_2, PaO_2, and $PaCO_2$. Their mean and 95% confidence interval is shown in Table 3.2 for patients in their 72h, and in "Appendix A" (Tables A4 and A5) at 24h and 48h, respectively. The main target variable was ARDS severity (where 0=mild, 1=moderate, and 2=severe). ICU mortality rates at 72h (Fig. 3.2), 24h (Figure A3 in Appendix A), and 48h (Figure A4 in Appendix A), and the duration of MV (Table A6 in Appendix A), were also obtained.

Table 3.2. Input variables and their descriptive statistics in eICU at 72-h according to PaO_2/FiO_2.

	Mild	Moderate	Severe	All
A. ARDS Patients	***872 (37.49%)***	***1,025 (44.07%)***	***429 (18.44%)***	***2,326 (100%)***
*B. Descriptive feature– **means and 95% CI***				
Age	64.77 [63.78, 65.76]	62.73 [61.83, 63.64]	59.97 [58.67, 61.28]	62.99 [62.39, 63.59]
PEEP	5.95 [5.80, 6.09]	7.16 [6.99, 7.34]	10.09 [9.72, 10.46]	7.25 [7.12, 7.38]
FiO_2	0.40 [0.39, 0.41]	0.50 [0.49, 0.51]	0.81 [0.79, 0.83]	0.52 [0.51, 0.53]
PaO_2	98.89 [97.17, 100.62]	80.52 [79.25, 81.78]	74.81 [72.83, 76.79]	86.36 [85.34, 87.37]
$PaCO_2$	39.93 [39.33, 40.53]	42.38 [41.72, 43.04]	44.23 [43.13, 45.33]	41.80 [41.38, 42.23]

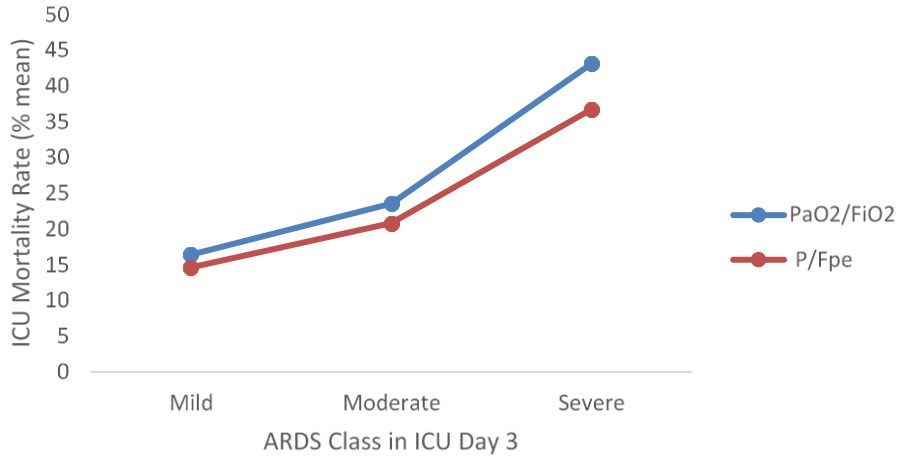


Fig. 3.2. Intensive Care Unit mortality rate at 72 h in relation to degree of lung severity in patients with acute respiratory distress syndrome included in the eICU database, according to PaO_2/FiO_2 ratio and P/FP_E .

3.2.2 Proposing P/FP_E and the New Mild, Moderate, and Severe Thresholds

Determination of ARDS severity continues to be a challenge. The current Definition of ARDS for Oxygenation is a PaO_2/FiO_2 ratio of 300 to 200 mmHg for mild, 200 to 100 for moderate and less than 100 for severe ARDS: for PEEP ≥ 5 cmH_2O. However, the empirical PaO_2/FiO_2 cut-offs for "severity" of 100, 200, and 300 mmHg are arbitrary and poorly validated [13]. Usually, arterial oxygenation in ARDS patients improves substantially by increasing the level of PEEP. Moreover, it has been clinically observed that increase in PEEP level with the same FiO_2 yields different PaO_2 and SpO_2 [41].

Besides, Berlin definition does not account for the nonlinear relationship of PaO_2 and FiO_2 [17] and has a limited predictive accuracy in recent trials [18-21]. Since the current Berlin definition for ARDS does not account for PEEP level in its computation, it provides an incomplete picture of the "*true*" ARDS severity. Accordingly, we propose a novel formula [$PaO_2/(FiO_2 x PEEP)$ or P/FP_E] for PEEP $\geq$ 5 cmH_2O to address Berlin's definition gap in calculating ARDS severity by incorporating PEEP in the new P/FP_E index to precisely determining ARDS severity, which could be better than the current definition.

Before starting our analysis, the thresholds of the P/FP_E index (for PEEP$\geq$5) were experimentally tuned. We computed the minimum and maximum P/FP_E values of the patients in the two databases, which were 2 and 60 $mmHg/cmH_2O$, respectively. Then, several cut-offs were studied in order to determine the ones that could be more accurate in the stratification of ARDS severity. For this purpose, we tested round values (to be easily remembered by intensivists) in the range 2-60, and analyzed P/FP_E index of the ARDS severity groups obtained. The partition showing a better separation of the ARDS severity groups obtained was achieved in this study, for the following intuitive classification thresholds (for PEEP $\geq$ 5 cmH_2O): 60-40 $mmHg/cmH_2O$ for mild, 40-20 for moderate, and <20 for severe, which are different from the Berlin definition, and within each of these ARDS categories, the patients had a similar degree of lung severity.

3.2.3. Testing the Benefits of P/FP_E Index with Respect to the Standard PaO_2/FiO_2 Ratio

Our study was based on ML analysis and not on the conventional statistical hypothesis testing analysis. In general, ML is an exploratory process and a current application of AI to generate predictive models. Using this technology, there is not a one-model-fits-all solution. Precisely, there is no ML method that reaches the highest accuracy for all

domains, datasets, or problem types [48]. The optimal model differs from one problem to another based on the characteristics of variables and observations. Since in a substantial proportion of patients diagnosed as having ARDS did not meet ARDS criteria within the first 24 hours of care, we aimed at implementing ML models capable of predicting ARDS severity over time to compare the PaO_2/FiO_2 ratio -as mandated by the current Berlin criteria for ARDS- with the proposed new P/FP_E ratio according to the following three scenarios: (i) Scenario I: predicting ARDS severity in the 3rd ICU day using information captured in the 1st ICU day; (ii) Scenario II: predicting ARDS severity in the 3rd ICU day using information captured in the 2nd ICU day; (iii) Scenario III: predicting ARDS severity in the 3rd ICU day using information captured in the 1st and 2nd ICU days.

We implemented three robust supervised ML predictive models using Python 3.7. These models were based on the ML algorithms RF [29], XGBoost [30], and LightGBM [31]. Grid search was used to identify the optimal values for their input parameters. The quality of the prediction models was computed based on a 10-fold cross-validation approach. AUC and CORR were used to assess model performance in predicting ARDS severity as a categorical prediction.

To provide a meaning to the findings, we used the classification of performance suggested by Hosmer and Lemeshow [49]: “excellent” if AUC ≥ 0.9; “good” if AUC is between 0.8 and 0.9; “fair” if AUC is between 0.7 and 0.8; “poor” if AUC is between 0.6 and 0.7; and “very poor” if AUC is below 0.6.

For CORR, we used the interpretation suggested by Mukaka [50] who proposed “very high” for CORR ≥0.9 (positive correlation) or CORR ≤-0.9 (negative correlation); “high” if CORR is between 0.7 and 0.9 (positive) or -0.9 and -0.7 (negative); “moderate” if CORR is between 0.5 and 0.7 (positive) or -0.7 and -0.5 (negative); “low” if CORR is between 0.3 and 0.5 (positive) or -0.5 and -0.3 (negative), and “negligible” otherwise.

3.3 Results

The findings of the three classification ML methods for the three predictive scenarios in the two databases are presented in Tables 3.3 and 3.4. Table 3.3 shows the quality of ML predictions for MIMIC-III, confronting the results obtained for PaO_2/FiO_2 (Table 3.3a) with those obtained for P/FP_E (Table 3.3b). Table 3.4 shows the same comparative results in patients from the eICU database.

Table 3.3. Quality of the third ICU day severity predictive ML models for MIMIC-III (means and standard deviations are shown)

(a) PaO_2/FiO_2 results

Scenario I: Predicting ARDS Severity in the 3rd ICU day using the data in 1st ICU day		
Algorithm	**AUC, mean ± SD**	**CORR, mean ± SD**
XGBoost	*0.616 ± 0.039*	*0.190 ± 0.068*
RF	*0.622 ± 0.048*	*0.173 ± 0.089*
LightGBM	*0.612 ± 0.039*	*0.138 ± 0.084*
***Scenario II: Predicting ARDS Severity in the 3rd ICU day using the data in 2nd ICU day**		
Algorithm	**AUC, mean ± SD**	**CORR, mean ± SD**
XGBoost	*0.621 ± 0.023*	*0.147 ± 0.121*
****RF***	***0.635 ± 0.020***	*0.139 ± 0.094*
LightGBM	*0.622 ± 0.025*	*0.126 ± 0.120*
Scenario III: Predicting ARDS Severity in the 3rd ICU day using the data in 1st & 2nd ICU days		
Algorithm	**AUC, mean ± SD**	**CORR, mean ± SD**
XGBoost	*0.619 ± 0.030*	*0.150 ± 0.106*
RF	*0.627 ± 0.022*	*0.177 ± 0.108*
LightGBM	*0.618 ± 0.022*	*0.086 ± 0.101*

***** Identifies the optimal scenario and ML model

(b) P/FP_E results

Scenario I: Predicting ARDS Severity in the 3rd ICU day using the data in 1st ICU day		
Algorithm	**AUC, mean ± SD**	**CORR, mean ± SD**
XGBoost	*0.711 ± 0.029*	*0.385 ± 0.064*
RF	*0.712 ± 0.027*	*0.408 ± 0.060*
LightGBM	*0.716 ± 0.029*	*0.376 ± 0.073*
***Scenario II: Predicting ARDS Severity in the 3rd ICU day using the data in 2nd ICU day**		
Algorithm	**AUC, mean ± SD**	**CORR, mean ± SD**
XGBoost	*0.785 ± 0.025*	*0.514 ± 0.053*
RF	*0.787 ± 0.023*	*0.546 ± 0.061*
****LightGBM***	***0.788 ± 0.020***	***0.566 ± 0.044***
Scenario III: Predicting ARDS Severity in the 3rd ICU day using the data in 1st & 2nd ICU days		
Algorithm	**AUC, mean ± SD**	**CORR, mean ± SD**
XGBoost	*0.782 ± 0.025*	*0.548 ± 0.049*
RF	*0.780 ± 0.023*	*0.538 ± 0.065*
LightGBM	*0.785 ± 0.021*	*0.511 ± 0.055*

Table 3.4. Quality of the third ICU day severity predictive ML models for eICU (means and standard deviations are shown)

(a) PaO_2/FiO_2 results

Scenario I: Predicting ARDS Severity in the 3rd ICU day using the data in 1st ICU day		
Algorithm	**AUC, mean ± SD**	**CORR, mean ± SD**
XGBoost	*0.712 ± 0.032*	*0.398 ± 0.061*
RF	*0.714 ± 0.030*	*0.393 ± 0.059*
LightGBM	*0.713 ± 0.028*	*0.373 ± 0.069*
***Scenario II: Predicting ARDS Severity in the 3rd ICU day using the data in 2nd ICU day**		
Algorithm	**AUC, mean ± SD**	**CORR, mean ± SD**
****XGBoost***	***0.863 ± 0.016***	***0.725 ± 0.028***
RF	*0.863 ± 0.016*	*0.700 ± 0.040*
LightGBM	*0.860 ± 0.014*	*0.714 ± 0.028*
Scenario III: Predicting ARDS Severity in the 3rd ICU day using the data in 1st & 2nd ICU days		
Algorithm	**AUC, mean ± SD**	**CORR, mean ± SD**
XGBoost	*0.860 ± 0.015*	*0.717 ± 0.025*
RF	*0.854 ± 0.017*	*0.693 ± 0.038*
LightGBM	*0.857 ± 0.014*	*0.713 ± 0.027*

(b) P/FP_E results

Scenario I: Predicting ARDS Severity in the 3rd ICU day using the data in 1st ICU day		
Algorithm	**AUC, mean ± SD**	**CORR, mean ± SD**
XGBoost	*0.735 ± 0.034*	*0.525 ± 0.056*
RF	*0.735 ± 0.034*	*0.514 ± 0.057*
LightGBM	*0.734 ± 0.034*	*0.511 ± 0.053*
***Scenario II: Predicting ARDS Severity in the 3rd ICU day using the data in 2nd ICU day**		
Algorithm	**AUC, mean ± SD**	**CORR, mean ± SD**
****XGBoost***	***0.873 ± 0.022***	***0.745 ± 0.033***
RF	*0.868 ± 0.016*	*0.739 ± 0.039*
LightGBM	*0.869 ± 0.023*	*0.728 ± 0.043*
Scenario III: Predicting ARDS Severity in the 3rd ICU day using the data in 1st & 2nd ICU days		
Algorithm	**AUC, mean ± SD**	**CORR, mean ± SD**
XGBoost	*0.872 ± 0.020*	*0.725 ± 0.040*
RF	*0.860 ± 0.015*	*0.731 ± 0.038*
LightGBM	*0.871 ± 0.022*	*0.717 ± 0.040*

For MIMIC-III, the best ML severity predictive model on the ARDS severity in the third ICU day was obtained by scenario II and by P/FP_E with an AUC = 0.788 and CORR = 0.566, using LightGBM algorithm. When PaO_2/FiO_2 was used, the values were AUC = 0.635 and CORR = 0.19, but these performances were obtained with different algorithms. In qualitative

terms, P/FP_E ratio improved PaO_2/FiO_2 ratio from "poor" to "fair" AUC, and from "negligible" to "moderate" CORR.

For the eICU database, the results were slightly better. The best ML severity predictive model was also observed for scenario II. This finding confirms that the best approach to predict ARDS severity on the third ICU day is to consider the condition of the patient in the second ICU day after ARDS onset, rather than the first ICU day or both. For eICU data, the best AUC and CORR values are 0.873 and 0.745 for P/FP_E; and 0.863 and 0.725 for PaO_2/FiO_2. These results are qualified as a "good" predictive accuracy and a "high" correlation.

In general, results show that P/FP_E ratio has a better behavior in the prediction of ARDS severity than PaO_2/FiO_2 ratio in terms of AUC and CORR. Whereas PaO_2/FiO_2 obtained up to 0.635 AUC and up to 0.19 CORR in MIMIC-III, the use of P/FP_E reached 0.788 AUC and 0.566 CORR. This represents increments of +0.153 AUC and +0.376 CORR and shows the advantages of using the P/FP_E ratio.

3.4 Discussion

In this large study, we proposed a novel formula (P/FP_E) and corresponding thresholds for classifying ARDS severity. We investigated several ML methods to generate severity predictive models in almost 8,000 patients with ARDS over time after ARDS diagnosis. Our findings confirmed that the best approach to predict ARDS severity on the third ICU day is to consider the condition of the patient in the second ICU day after ARDS onset, rather than during the first ICU day as mandated by Berlin criteria.

For the MIMIC-III database, predictive models using the P/FP_E ratio attained outstanding improvements in terms of AUC (15% improvement) and CORR (37.6% improvement), when compared to the previous PaO_2/FiO_2 models. For the eICU database, models based on P/FP_E also outperformed PaO_2/FiO_2 predictions, with 14.8% and 2% improvements of AUC and

CORR, respectively. The difference in terms of the accuracy between the two databases is remarkable regarding CORR. This is due to the fact that eICU is a multicenter ICU database with high granularity data (i.e., high level of detail in the data) for over 200,000 admissions to ICUs. By contrast, MIMIC-III is a single-center ICU database for approximately 60,000 admissions of ICU patients. Therefore, in all extracted data of the three ICU days, the number of extracted patients from eICU was greater than the number of extracted patients from MIMIC-III. Consequently, this would lead to better ML results in terms of CORR for the eICU database. Overall, the novel P/FP_E ratio outperformed the PaO_2/FiO_2 ratio in all ML applied models, and showed that predictions based on the patient condition in the second day after onset are better than predictions based on the first 24h (7.2-13.8% AUC and 1.5-22% CORR improvements), followed by the predictions based on both the first and the second day conditions (0.1-0.3% AUC and 0.18-14% CORR improvements).

In contrast to our study, most recent studies developed ML approaches to predict the risk of ARDS in critically ill patients prior to ARDS onset [46,51,52], based on single-center databases [46,51] and using one single ML algorithm [46,50]. Consequently, their findings have serious limitations for the generalizability in the context of assessing the prediction of ARDS outcome.

This large study proposes a novel criterion to reclassify ARDS patients in terms of severity by using ML methods on an extensive amount of data from two large datasets of critically ill patients. The relatively good accuracy of P/FP_E (when compared to PaO_2/FiO_2) in stratifying ARDS patients could allow to overcome the major clinical drawbacks of the current Berlin definition. Also, this study is implementing ML models for predicting severity over time after ARDS onset. Critically ill patients are an ideal population for clinical database investigations using machine learning algorithms because while the data from ICUs are extensive, the value of many diagnostic and therapeutic interventions remains largely unproven [53].

ARDS is considered one of the major reasons of ICU admission and it is associated with a high hospital mortality [1]. Despite its high mortality rate and high rates of ICU utilization, ARDS remains critically misdiagnosed and globally under-diagnosed in the ICU settings [1]. Furthermore, increasing ARDS severity is associated with increased mortality rate [6]. The PaO_2/FiO_2 ratio categorizes ARDS patients according to the severity of their oxygenation deficit without considering the level of applied PEEP in the assessment of lung severity. The PaO_2/FiO_2 ratio does not appropriately show the severity of ARDS for PEEP≥5. However, the application of PEEP plays a significant role in improving oxygenation. It is well established that changes in PEEP alter the PaO_2/FiO_2 in lung-injured patients [41]. Attempting to predict lung severity and patient outcomes based solely in PaO_2/FiO_2 on this basis is inherent flawed. Thus, the stratification of ARDS patients as proposed by the Berlin criteria is useless for assessing severity of lung injury and could be of no benefit for enrolling patients into therapeutic clinical trials. Our P/FP_E formula for PEEP≥5, appropriately addressed Berlin's definition gap in computing ARDS severity by including PEEP in the novel ratio. Clearly, our study showed that P/FP_E thresholds improved prediction of ARDS severity. This can lead to important medical implications by accurately anticipate specific treatment for each ARDS category, which could eventually decrease ARDS mortality. In other words, P/FP_E can represent a good solution for the clinical assessment of ARDS severity and as a guidance for treatment of ARDS.

Our study has several strengths. First, predicting ARDS severity using ML algorithms is feasible. Our ML study overcomes several challenges in predicting ARDS severity, including issues with data and the heterogeneity of operationalizing ARDS severity as an outcome label, model development issues, and generalizability. Second, we have analyzed a large population of ARDS patients within their first three ICU days after onset. Third, we have described and validated our findings using both a large single-center database (MIMIC-III) and a large

multicenter database (eICU). Forth, we have investigated several ML predictive models for ARDS severity over time after ARDS onset. We believe that our approach is generalizable across other ARDS populations. However, we acknowledge some limitations to our study. First, our work is based on a retrospective analysis of data whose results concerning P/FP_E benefits should be confirmed in further prospective studies. Second, our analysis is concerned with the evolution and stratification of patients in their third ICU day after ARDS onset. Although the first 72-h are essential in the management and progression of ARDS patients, our study lacks the assessment of a long-term outcome (e.g., ICU mortality, 60-day mortality). Third, further longitudinal studies on complete evolution of ARDS patients could help to find out new evidence(s) on the management of ARDS since our ML results achieved outstanding improvements compared to the current state, with “fair” to “good” predictions of ARDS severity [49]. Forth, one could argue that extracorporeal membrane oxygenation (ECMO) is not considered in this study. ECMO is a clinical outcome and can only temporarily sustain severe ARDS patients to bridge periods of time when oxygenation through the lungs cannot be achieved via MV. Moreover, ECMO is a constrained resource that is not available in all ICUs. Hence, for the purpose of our study, we only considered patients receiving MV for >48 hours [42,43]. Fifth, regarding the potential consequences of using the new ratio at the bedside, further studies are needed to examine whether it could help for clinical decision making and guiding therapy. Our study opens a possibility to better define ARDS severity, as a new research area for patient care improvement.

3.5 Conclusions

This large study proposes a novel criterion based on the P/FP_E formula to assess ARDS severity using ML, which is significantly better than the current Berlin criteria using baseline PaO_2/FiO_2. We are conscious that, from a technical point of view, the AUC and CORR

improvements are moderate but, from a clinical point of view, these improvements are significantly relevant. Concretely, applying the proposed new criteria for ARDS severity enables critical care physicians to assess lung severity by involving PEEP information. Moreover, being able to better adjust the severity profiles of ARDS patients will potentially improve the selection of more adequate therapeutic regimens for each ARDS category, which could contribute to reduce ARDS mortality. However, additional studies are required in order to confirm this. In both databases (MIMIC-III and eICU) and either in Berlin or P/FP_E, scenario II (assessment of oxygenation deficit after 24 h of ARDS diagnosis and routine ICU treatment) was the best severity predictive scenario. From a ML perspective, P/FP_E outperformed PaO_2/FiO_2 in all ML models predicting ARDS severity after onset over time in all scenarios either in MIMIC-III or eICU. Accordingly, this study can serve as an example of how ML is a worth-considering technology to gain new insights in the development of ARDS predictive models which could contribute to improve ICU resource allocation and mortality reduction.

CHAPTER 4

Predicting Duration of MV in ARDS Using ML

4.1 Introduction

A recent study argues that mild ARDS should be considered "severe in terms of level of care" [35]. This quality criterion (i.e., level of care) could be measured in terms of MV duration, but accurate predictions of MV duration are hard for critical care physicians [54,55], particularly for patients requiring prolonged MV [55].

Predicting MV duration influences important clinical decisions, such as timing of tracheostomy and initiation of oral nutrition [55]. In this context, one approach for an accurate prediction of MV duration could be done with AI technologies, such as ML. ML has been used in ICU to predict clinical outcomes [56-62]. Troché and Moine addressed the critical question on whether MV duration is predictable [61] reaching a conclusion that prediction of MV duration is a difficult task.

In this chapter, we present the use of three powerful supervised ML methods to develop novel models to predict MV duration in ARDS after onset over time using the single-center MIMIC-III dataset under three different scenarios. Then, the eICU multicenter dataset is used to externally validate the best MIMIC-III prediction model. Consequently, this chapter addresses the second objective of this PhD thesis: proposing a ML-based model for predicting duration of MV in ARDS.

4.2 Methods

4.2.1 Study Design and Patient Populations

As in the previous study described in chapter 3, we base our study in the datasets MIMIC-III [32] and eICU [33], but here we used these two publicly available clinical datasets for development and external validation of the best ML predictive model, respectively. The selection of clinical variables was based on prior studies [19,46,59,63,64]. All extracted patients from both datasets fulfilled the Berlin definition for ARDS [6]. For the purpose of this study, prolonged MV was defined as being ventilated for >48 h [62,65]. Disease progression in each dataset was tracked along those 3 ICU days.

MIMIC-III

The predictors were six: baseline demographic information (age); the ventilator parameters PEEP; and blood gas parameters including FiO_2, PaO_2, PaO_2/FiO_2, and $PaCO_2$. The main target variable was MV duration.

eICU

We used this dataset for external validation of the best prediction model obtained from MIMIC-III dataset, in order to obtain the MV duration prediction in the eICU dataset. The clinical parameters extracted from eICU dataset were the same that we took for MIMIC-III dataset.

4.2.2 Predictive Models for MV Duration of ARDS Patients

During the first 24 hours of ARDS onset, misdiagnosis can occur if clinicians consider qualifying PaO_2 values resulting from acute events unrelated to the disease process (such as endotracheal tube obstruction, barotrauma, or hemodynamic instability), instead of considering only PaO_2 values while patients are clinically stable. It is also well established that changes in

PEEP and FiO_2 within the first few hours of intensive care management alter the PaO_2/ FiO_2 ratio in ARDS patients [8].

Since in a substantial proportion of patients diagnosed as having ARDS did not meet ARDS criteria within the first 24 hours of care, we decided to examine supervised ML models in the following three scenarios during the first two ICU days: (i) Scenario I: predicting MV duration using information captured in the 1st ICU day; (ii) Scenario II: predicting MV duration using information captured in the 2nd ICU day; (iii) Scenario III: predicting MV duration using information captured in the 1st and 2nd ICU days- Then we compared these three scenarios with scenario IV for predicting MV duration using the information captured in the 3rd ICU day exclusively.

We implemented three robust supervised ML algorithms via Python 3.7 using LightGBM [31], RF [29], and XGBoost [30] to generate predictive models for MV duration after ARDS onset over time in the development dataset (i.e., MIMIC-III). For external validation purposes, we used the multicenter eICU dataset as these three algorithms sacrifice the explicitness of the model in favor of predictive quality, and the generated models should be seen as "black boxes" with a high predictive robustness.

For the development dataset, we optimized each model's parameters through a grid search over the respective model's hyperparameter space and the quality of all prediction models was computed based on a 10-fold cross-validation approach.

RMSE was used to assess the predictive quality of the models. RMSE quantifies more significant differences between the predicted and the actual patient readings when they occur [66]. MV duration was expressed in days.

4.3 Results

For the development dataset (i.e., MIMIC-III) and the validation dataset (i.e., eICU), the mean values and the 95% confidence intervals (CI) of the baseline parameters during the first three ICU days after ARDS onset are reported in Table 4.1. The median and interquartile range (IQR) of MV duration are also reported in Table 4.2.

Table 4.1. Predictors and their descriptive statistics in MIMIC-III and eICU at 24-h, 48-h, and 72-h

	24-h	48-h	72-h
A. MIMIC-III ARDS Patients	***2,466 (100%)***	***1,445 (58.6%)***	***1,278 (51.8%)***
B. Means and 95% CI			
Age	62.2 [61.5, 62.8]	60.8 [59.9, 61.6]	60.9 [60.0, 61.8]
PEEP	7.6 [7.5, 7.7]	9.1 [8.9, 9.4]	8.9 [8.8, 9.2]
FiO_2	0.66 [0.65, 0.67]	0.54 [0.53, 0.55]	0.51 [0.49, 0.51]
PaO_2	114.5 [112.8, 116.2]	97.6 [96.3, 98.9]	95.4 [94.1, 96.6]
$PaCO_2$	43.4 [42.9, 43.9]	42.3 [41.8, 42.9]	42.9 [42.4, 43.6]
PaO_2/FiO_2	184.3 [181.9, 186.6]	170.9 [167.7, 174.2]	179.1 [175.7, 182.5]
C. eICU ARDS Patients	***5,153 (100%)***	***2,981 (57.8%)***	***2,326 (45.1%)***
D. Means and 95% CI			
Age	63.4 [62.9, 63.8]	63.4 [62.8, 63.9]	62.9 [62.4, 63.6]
PEEP	6.6 [6.6, 6.7]	7.1 [7.0, 7.2]	7.3 [7.1, 7.4]
FiO_2	0.63 [0.63, 0.64]	0.53 [0.52, 0.54]	0.52 [0.51, 0.53]
PaO_2	104.1 [102.9, 105.2]	89.1 [88.1, 90.1]	86.4 [85.3, 87.4]
$PaCO_2$	43.5 [43.2, 43.9]	41.3 [40.9, 41.7]	41.8 [41.4, 42.2]
PaO_2/FiO_2	160.2 [158.3, 162.1]	175.2 [172.9, 177.5]	174.5 [171.8, 177.2]

Table 4.2. MV Duration in ARDS across MIMIC-III and eICU.

ICU Day (*N*)	Database	MV duration *Median days (IQR days)*
Day 1 (*2,466*)	*MIMIC-III*	6.5 (4.4–9.8)
Day 2 (*1,445*)		6.8 (4.7–10.5)
Day 3 (*1,278*)		6.9 (4.7–10.6)
Day 1 (5,153)	*eICU*	5.0 (3.0–9.0)
Day 2 (2,981)		6.0 (4.0–10.0)
Day 3 (2,326)		6.0 (4.0–10.0)

Table 4.3 shows the performance of the three supervised ML methods for the three predictive scenarios when trained with the data in the development database, and subject to 10-fold cross-validation. Table 4.4 shows the results of the external validation of the best prediction model obtained from MIMIC-III when confronted to the prediction of the MV duration in the eICU dataset. All performance values are expressed as RMSE mean ± standard deviation values. Best results are highlighted in bold.

Table 4.3. Performances of LightGBM, RF, and XGBoost models to predict MV duration over time in MIMIC-III.

Scenario I: Predicting MV duration in ARDS using data in the 1st ICU day	
Algorithm	**RMSE, mean ± SD**
XGBoost	*6.81 ± 1.18*
RF	*6.79 ± 1.22*
LightGBM	***6.41 ± 1.55***
***Scenario II: Predicting MV duration in ARDS using data in the 2nd ICU day**	
Algorithm	**RMSE, mean ± SD**
XGBoost	*6.53 ± 0.96*
RF	*6.55 ± 1.16*
****LightGBM***	***6.10 ± 0.72***
Scenario III: Predicting MV duration in ARDS using data in the 1st & 2nd ICU days	
Algorithm	**RMSE, mean ± SD**
XGBoost	*6.57 ± 1.08*
RF	*6.60 ± 1.01*
LightGBM	***6.35 ± 0.69***
Scenario IV: Predicting MV duration in ARDS using the data in the 3rd ICU day	
Algorithm	**RMSE, mean ± SD**
XGBoost	*6.14 ± 0.85*
RF	*6.19 ± 0.66*
LightGBM	***5.92 ± 0.47***

***** Identifies the optimal scenario and ML model

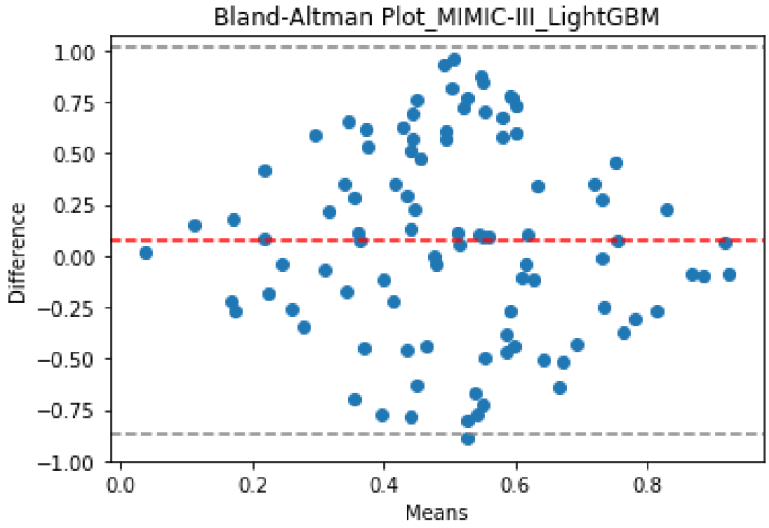
Table 4.4. External validation of the best prediction model (LightGBM) obtained from MIMIC-III to obtain the MV duration prediction in the eICU database.

Predictive Scenario	RMSE, mean ± SD
Scenario I	*6.08 ± 0.72*
***Scenario II**	***5.87 ± 0.67***
Scenario III	*5.93 ± 0.44*
Scenario IV	*5.71 ± 0.55*

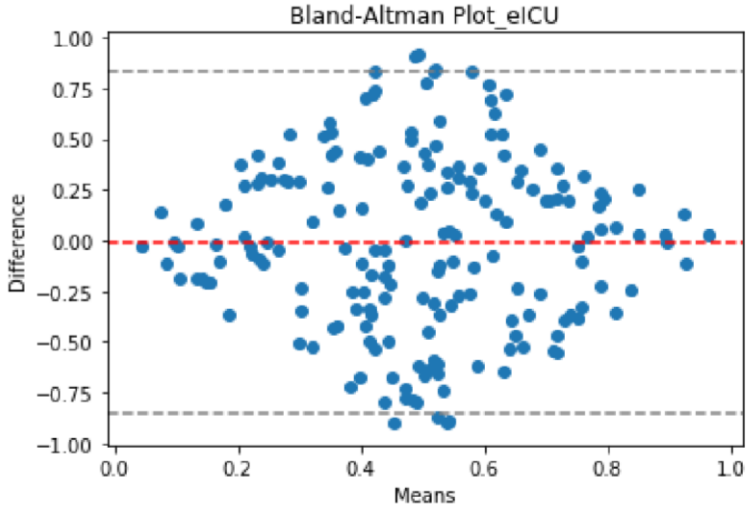
For the development database, the best early ML model for predicting MV duration was obtained by scenario II, with an average RMSE = 6.10 days, using the LightGBM algorithm.

For the validation database, the best early ML predictive model for MV duration was also observed for scenario II with an average RMSE = 5.87 days. This finding reinforces the idea that the best early approach for predicting MV duration is to consider the condition of the patient in the second ICU day after ARDS onset, rather than the first ICU day, or both.

Bland-Altman plots shows how much similar new instruments or techniques (e.g., the LightGBM model) are at measuring something (e.g., the MV duration) in comparison to the instruments or techniques currently being used (i.e., the real prediction in scenario II). The X- and Y-axes stand for the mean and the difference of the two measurements, respectively. As shown in Figures 4.1a and 4.1b, differences between predicted and real values showed a relatively spread around 0 across the range of MV duration average. In other words, being close to zero means that LightGBM model predictions are very close to real values either for the MIMIC-III or the eICU datasets. In general, the Bland-Altman plots illustrated agreement between the LightGBM models and real MV durations using the development and validation datasets.



(a)



(b)

Fig. 4.1. Bland–Altman plot for the truth vs. the predicted values of MV duration using LightGBM (the best validated model) in Scenario II (the best early scenario). (a) Development Database; (b) Validation Database. The X- and Y-axes stand for the mean and the difference of the two measurements, respectively. Please note that the values shown in the Bland-Altman plot are normalized in the interval [0,1] (i.e., values are scaled to have corresponding values between 0 and 1).

4.4 Discussion

Comparing the difference of RMSE means in the best early scenario (scenario II) with the prediction based on the data of patients in their third ICU day (scenario IV), yields minor

RMSE differences [development database: 0.18 day (6.10–5.92)] for LightGBM and [validation database: 0.16 day (5.87–5.71)]. According to these low differences for both the development and validation datasets, our major finding was that the prediction results of LightGBM models based on the data of the second ICU day (scenario II) are very close to those corresponding results of LightGBM models based on the data of the third ICU day (scenario IV). Consequently, the LightGBM model can accurately predict MV duration without considering/waiting for the data of the third ICU day. This means that MV duration can be predicted earlier, and this will lead to better allocation of MV resources, reducing high acute costs of MV in ARDS, and improving patient care.

MV duration beyond 48 hours in patients with ARDS provides information about risk factors in those patients [65] and has a direct correlation with ICU costs [25,26]. An early prediction of MV duration can optimize ICU-level resource utilization [26,67]. Previous attempts to predict MV duration using conventional ICU scores, or traditional statistical regression-based techniques have proven to be difficult and failed to deal with the diversity of big data in the modern ICU databases [62]. ML has a reliability, and it is a non-invasive modality to generate models for effectively predicting MV duration. Most previous works considered a discriminative prediction model to determine if a patient will remain intubated after a fixed number of days (e.g., 7 days) [62]. By contrast, our approach is continuous numerical, and it early predicts the number of MV days using commonly accessible clinical variables during the first two ICU days. Furthermore, to strengthen the evidence of our results, we used a multicenter database (eICU) for external validation, in which the best model obtained from a single-center database (MIMIC-III) was used to obtain the MV duration prediction in the eICU database.

Our findings could be used to facilitate optimal triage, more timely management, and ICU resource utilization [68]. They may also affect some important clinical decisions, including

timing of tracheostomy and potentially transfers to long-term ventilator weaning units or referral to other centers [54].

Herein, the main objective of using ML was to show that the application of ML is a promising approach to early predict MV duration. The ML contribution in this large study is to demonstrate the applicability of this approach, while not trying to choose the most proper ML model. Furthermore, we believe that the results of an efficient ML technique can yield accurate results for predicting MV duration. In terms of clinical relevance, our ML findings showed that using clinical data from the first ICU day is less predictive than data from the second ICU day. Previous studies showed that the accuracy of intensivists to predict MV duration is limited [54]. Although comparing our results with other published ML predictions of MV duration is difficult, as we aimed at predicting MV duration for MV >48 h. On the contrary, predictions of prior studies are for different outcomes and under different time frames, in different populations, and using different ML metrics. A recent ML study showed that RMSE for predicting MV duration in ARDS patients for MV >48 h, was 6.23 days [63]. However, this study in [63] has several weaknesses: (1) it ignores the temporal dependency of the longitudinal predictor and treats each observed data point independently and (2) it is only based on the single-center MIMIC-III database without an external validation. Hence, those findings have serious limitations for generalizability in the context of assessing the prediction of ARDS outcome.

From a cost perspective, the mean incremental cost of MV in ICU patients in the US was $1,522 per day [25]. For instance, if we compare our findings with the result of the best ML method used in [63], which had a RMSE of 6.23 days, we see that LightGBM approach (the best approach) improved the current state of the art. This improvement can be quantified in terms 0.13 day (6.23–6.10) and about US $198 per patient according to [25]. Developing early predictive models using ML could assist to implement policies for the reduction of high acute

care costs in ARDS [24-26]. Previous clinical studies showed acute costs incurred by mechanically ventilated ICU patients, but there is a significant difference in costs between ventilated ARDS patients and those without ARDS [69]. More specifically, ARDS diagnosis increases total ICU and hospital costs for mechanically ventilated ICU patients, suggesting higher total costs due to more days on a ventilator, although there is no clear severity-dependent relationship between ARDS severity and incurred costs [69]. The benchmarking of ML algorithms is possible through publicly available databases such as MIMIC-III [59,46] or eICU [59,70].

We acknowledge that our study has several strengths. First, the use of ML algorithms to construct MV duration prediction models overcomes the natural limitation of intensivists and medical approaches to obtain good predictions. Second, we have analyzed a large population over 7,000 ARDS patients from two ICU databases within the first three ICU days after ARDS onset. Third, we have implemented and externally validated the best ML model (LightGBM) that can accurately and early predict MV duration using commonly accessible clinical variables. Forth, early prediction of MV duration can inform population-level ICU resource allocation. Despite its strengths, we also acknowledge some limitations. First, our study is based on a retrospective analysis of data and should be confirmed through further prospective studies. Second, one could argue that the outcome of MV duration is somewhat subjective and could be a function of local practice or intrinsic bias inherent in such critical care decisions. However, our ability to early predict a clinically relevant and hard outcome (MV duration) supports the value of the proposed supervised ML models.

4.5 Conclusions

Predicting MV duration after ARDS onset over time is complex and cannot be adequately performed by critical care physicians, clinical scales, or medical technologies. Our findings

showed that the ML-based early prediction of MV duration is more accurate when predictive models are based on the clinical features of ARDS patients in the second ICU day after ARDS onset.

CHAPTER 5

Validating P/FP_E Index to Predict Mortality in ARDS Using ML

5.1 Introduction

ARDS is a noncardiogenic pulmonary edema, lung inflammation with hypoxemia, and decreased lung compliance. ARDS is a heterogeneous syndrome with a fatal outcome, a constellation of clinical and physiologic observations thought to represent a common pathology. Pathogenesis of ARDS remains elusive, and there is no gold standard diagnostic test. There is a lot of heterogeneity in ARDS diagnosis, the possibility that ARDS is, in fact, a collection of different diseases that have not yet been separately identified [1].

The panel of experts of the Berlin definition stratified lung severity of ARDS patients into three categories (mild, moderate, and severe) based on the PaO_2/FiO_2 ratio at the time of ARDS onset or diagnosis, independent of the level of FiO_2 and applied PEEP (a minimum of 5 cmH_2O was required) [6]. Mortality in ARDS increases with disease severity. In the Large Observational Study to Understand the Global Impact of Severe Acute Respiratory failure (LUNG SAFE) study, patients with ARDS stratified by the Berlin definition were reported to have an unadjusted mortality of 35%, 40% and 46% in mild, moderate and severe ARDS, respectively [1]. In addition, the disease trajectory of patients within each ARDS category can impact outcome. When patients with initial mild ARDS in the LUNG SAFE study were sub-classified into three groups based on the evolution of severity in the first week as “worsening”,

"persisting", and "improving"; the mortality was reported to be 10%, 30% and 37% for patients with improving, persisting, and worsening ARDS, respectively [36]. Factors that influence ARDS outcomes can be related to the patient such as age and presence of comorbidities [71], to the treatments received such as positive fluid balance and packed red cell transfusions [72, 73], to the overall severity of illness as measured by scores such as the acute physiologic and chronic health evaluation (APACHE) [73], and to the assessment of respiratory parameters including gas exchange (ex. PaO_2/FiO_2 ratio and oxygen saturation index) [74] and dead space ventilation (ex. Ventilatory ratio) [75]. Hence, predicting the outcome of ARDS remains challenging and no scoring system has been validated until recently [76, 77].

Although the PaO_2/FiO_2 at ARDS onset is the most common criterion for assessing ARDS severity, it does not provide an accurate assessment for severity and outcome [59,76,78]. In addition, conventional ICU severity indices such as the simplified acute physiology score (SAPS-II) and sequential organ failure assessment (SOFA) can also be used to predict ICU outcome of ARDS patients. However, utilization of these severity indices is controversial since the factors impacting mortality in ARDS are multifaceted [76,78], and it is difficult for these indices to accurately predict mortality as they are generally a linear combination of explanatory variables and their generalizability in different ARDS cohorts may be limited [74,76,79]. Hence, none of them have been widely accepted for ICU mortality prediction in ARDS. Furthermore, previous attempts to predict ICU mortality using traditional statistical regression-based techniques have proven to be difficult and failed to deal with the diversity of big data in the modern ICU databases [60]. A more robust prediction system is urgently required.

Sayed et al. [59] proposed a novel criterion [$PaO_2/(FiO_2 x PEEP)$ or P/FP_E index] for PEEP≥5 to assess ARDS severity. The thresholds were 60 to 40 $mmHg/cmH_2O$ for mild, 40 to 20 for moderate, and less than 20 for severe ARDS. This new criterion addressed Berlin's definition gap in computing severity by incorporating the level of applied PEEP in the

oxygenation ratio. The P/FP_E index was markedly better than current PaO_2/FiO_2 ratio [59] in assessing ARDS severity after onset over time.

Several studies proved the applicability of machine learning (ML) in detecting adverse health events in the ICU settings [80,81]. However, no prior studies have investigated ICU mortality prediction ML-based models for each severity grade of ARDS over time, particularly beyond two ICU days.

It is in this context that we define the third of the objectives in this thesis: assessing the efficacy of the proposed P/FP_E index in comparison to the PaO_2/FiO_2 ratio with ML techniques for mortality prediction.

In this chapter, we conducted a derivation and validation study using a secondary analysis of two large, publicly available datasets of 7,619 ARDS patients to investigate the association between P/FP_E with its intuitive classification thresholds (different from the Berlin definition) vs. PaO_2/FiO_2 with ICU outcome, and examined the ML predictive performance of both indices for ICU mortality for each ARDS severity grade over time.

5.2 Methods

5.2.1 Study Design and Patient Populations

As in the previous study described in chapter 4, we used the MIMIC-III and eICU datasets, but here we used these publicly available ICU databases for the development and external validation of ML predictive models in the three severity grades of ARDS, separately. As it was mentioned before these databases correspond to a single-center MIMIC-III critical care dataset (MetaVision, 2008-2012) [32] and to a multicenter eICU dataset across the United States between 2014 and 2015 [33], respectively. Clinical variables were selected based on prior studies [59,83-86]. All the extracted patients from both datasets fulfilled the PaO_2/FiO_2 ratio [6] and the corresponding P/FP_E index [59]. For the purpose of this study, we only included

patients receiving MV for >48 hours [59,82]. Patients less than 18 years of age were excluded. Disease progression in each dataset was tracked along those 3 ICU days.

MIMIC-III

Predictors included baseline demographic information (age); gas exchange using blood gases parameters including, PaO_2, PaO_2/FiO_2 or P/FP_E, and $PaCO_2$; corresponding ventilator settings parameters including PEEP, FiO_2 (Table 5.1). The main target variable was ICU mortality for each ARDS severity grade (where 0=survival and 1=death) (Table 5.2).

eICU

We used this dataset for external validation of the prediction models of ICU mortality for each ARDS severity grade obtained from MIMIC-III (Tables 5.1 and 5.2).

Table 5.1. Predictors and their descriptive statistics in MIMIC-III and eICU at 24-h, 48-h, and 72-h.

	24-h	48-h	72-h
A. MIMIC-III ARDS Patients	***2,466 (100%)***	***1,445 (58.6%)***	***1,278 (51.8%)***
B. Means and 95% CI			
Age	62.2 [61.5, 62.8]	60.8 [59.9, 61.6]	60.9 [60.0, 61.8]
PEEP	7.6 [7.5, 7.7]	9.1 [8.9, 9.4]	8.9 [8.8, 9.2]
FiO_2	0.66 [0.65, 0.67]	0.54 [0.53, 0.55]	0.51 [0.49, 0.51]
PaO_2	114.5 [112.8, 116.2]	97.6 [96.3, 98.9]	95.4 [94.1, 96.6]
$PaCO_2$	43.4 [42.9, 43.9]	42.3 [41.8, 42.9]	42.9 [42.4, 43.6]
PaO_2/FiO_2	184.3 [181.9, 186.6]	170.9 [167.7, 174.2]	179.1 [175.7, 182.5]
P/FP_E	29.5 [28.9, 30.1]	24.2 [23.4, 25.0]	25.3 [24.5, 26.1]
C. eICU ARDS Patients	***5,153 (100%)***	***2,981 (57.8%)***	***2,326 (45.1%)***
D. Means and 95% CI			
Age	63.4 [62.9, 63.8]	63.4 [62.8, 63.9]	62.9 [62.4, 63.6]
PEEP	6.6 [6.6, 6.7]	7.1 [7.0, 7.2]	7.3 [7.1, 7.4]
FiO_2	0.63 [0.63, 0.64]	0.53 [0.52, 0.54]	0.52 [0.51, 0.53]
PaO_2	104.1 [102.9, 105.2]	89.1 [88.1, 90.1]	86.4 [85.3, 87.4]
$PaCO_2$	43.5 [43.2, 43.9]	41.3 [40.9, 41.7]	41.8 [41.4, 42.2]
PaO_2/FiO_2	160.2 [158.3, 162.1]	175.2 [172.9, 177.5]	174.5 [171.8, 177.2]
P/FP_E	28.2 [27.8, 28.6]	29.4 [28.9, 29.9]	29.2 [28.6, 29.9]

Table 5.2. ICU Mortality Rate distribution within ARDS classes over time across MIMIC-III and eICU.

Database	ICU Day (*N*)	Criterion	ARDS Class	ICU Mortality Rate (%) (*Mean and 95% CI*)
MIMIC-III	Day 1 (*718*)	PaO_2/FiO_2 Ratio	*Mild*	*22.4 [19.4, 25.5]*
	Day 2 (*472*)			*23.3 [19.5, 27.1]*
	Day 3 (*465*)			*22.8 [18.9, 26. 6]*
	Day 1 (*1,113*)		*Moderate*	*26.3 [23.7, 28.8]*
	Day 2 (*753*)			*27.6 [24.4, 30.8]*
	Day 3 (*661*)			*30.1 [26.6, 33.6]*
	Day 1 (*635*)		*Severe*	*32.5 [28.8, 36.1]*
	Day 2 (*220*)			*39.1 [32.6, 45.6]*
	Day 3 (*152*)			*40.8 [32.9, 48.7]*
	Day 1 (*711*)	P/FP_E Index	*Mild*	*23.6 [20.5, 26.8]*
	Day 2 (*280*)			*26.4 [21.2, 31.6]*
	Day 3 (*260*)			*23.1 [17.9, 28.2]*
	Day 1 (*927*)		*Moderate*	*26.5 [23.7, 29.4]*
	Day 2 (*459*)			*26.4 [22.3, 30.4]*
	Day 3 (*434*)			*29.0 [24.7, 33.3]*
	Day 1 (*828*)		*Severe*	*29.6 [26.5, 32.7]*
	Day 2 (*706*)			*29.6 [26.2, 32.9]*
	Day 3 (*584*)			*30.9 [27.2, 34.8]*
eICU	Day 1 (*1,549*)	PaO_2/FiO_2 Ratio	*Mild*	*14.3 [12.51, 16.0]*
	Day 2 (*1,098*)			*16.2 [14.0, 18.4]*
	Day 3 (*872*)			*16.4 [13.9, 18.9]*
	Day 1 (*2,361*)		*Moderate*	*18.7 [17.1, 20.3]*
	Day 2 (*1,454*)			*23.5 [21. 3, 25.7]*
	Day 3 (*1,025*)			*23.5 [20.9, 26.1]*
	Day 1 (*1,243*)		*Severe*	*28.7 [26.2, 31.3]*
	Day 2 (*429*)			*43.1 [38.4, 47.8]*
	Day 3 (*429*)			*43.1 [38.4, 47.8]*
	Day 1 (*1,309*)	P/FP_E Index	*Mild*	*14.4 [12.5, 16.3]*
	Day 2 (*849*)			*15.3 [12.9, 17.7]*
	Day 3 (*698*)			*14.6 [11.9, 17.2]*
	Day 1 (*2,053*)		*Moderate*	*18.1 [16.5, 19.8]*
	Day 2 (*1,182*)			*20.7 [18.4, 23.0]*
	Day 3 (*822*)			*20.8 [18.0, 23.6]*
	Day 1 (*1,791*)		*Severe*	*25.6 [23.6, 27.7]*
	Day 2 (*950*)			*34.7 [31.7, 37.8]*
	Day 3 (*806*)			*36.7 [33.4, 40.1]*

5.2.2 Predictive models

The current definition based on time of onset of ARDS does not accurately assess disease severity and does not provide reliable predictions of mortality. In addition, the disparate clinical

trajectory, and the challenge of heterogeneity in ARDS is daunting. Therefore, subsequent re-categorization of ARDS severity at fixed intervals (e.g., 24–72 h after disease onset) can provide an opportunity to classify more homogeneous subpopulations of patients in terms of disease progression and mortality [36,59,86]. We decided to examine ML models for predicting ICU mortality using the PaO_2/FiO_2 ratio vs. the P/FP_E index for each ARDS severity grade through the following three events: (i) Event I: ICU mortality prediction based on data from the 1st ICU day only; (ii) Event II: ICU mortality prediction based on data from the 2nd ICU day only; (iii) Event III: ICU mortality prediction based on data from the 3rd ICU day only. Figure 5.1 represents a workflow of the proposed methodology.

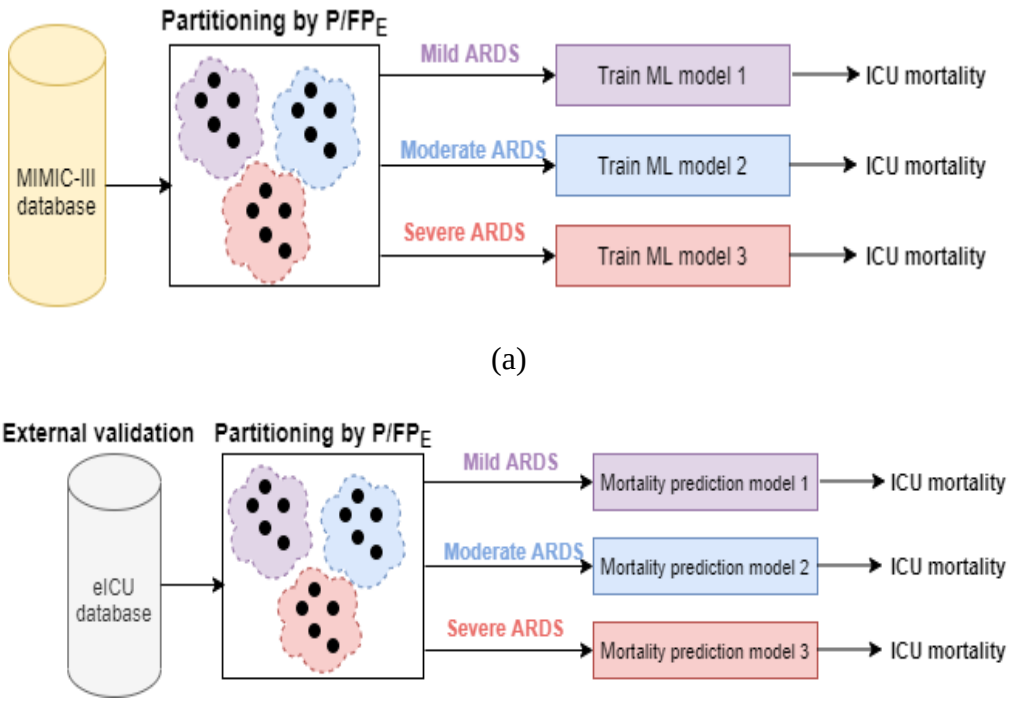


Fig. 5.1. Conceptual framework of the proposed approach. (a) Training the ML-based ICU mortality prediction model in the Development Dataset MIMIC-III; (b) External validation for the ML-based ICU mortality prediction model in the Validation Dataset eICU.

An algorithm using Random Forest [29] was implemented using Python 3.7 to generate predictive models for ICU mortality for each severity grade of ARDS over time in the

development dataset MIMIC-III. For external validation purposes, we used the multicenter eICU dataset. For the development dataset, the training parameters were optimized through a grid search over the respective model's hyperparameter space. The quality of all the prediction models was computed based on a 10-fold cross-validation. The AUC was used to assess the predictive performance of the ML models.

5.3 Results

The findings of the ML models for the three predictive events in the development and validation datasets are presented in Table 5.3 and 5.4. Table 5.3 shows the performance in terms of the mean and standard deviation of the AUC of ML predictions for the PaO_2/FiO_2 ratio, confronting the results obtained for the development dataset with those obtained for the validation dataset. Table 5.4 shows the comparative results for the P/FP_E index in the development and validation datasets, also expressed as the mean and standard deviation of the AUC.

Table 5.3. Performance of RF models: PaO_2/FiO_2 results

Event I: ICU mortality prediction in ARDS based on data of the 1st ICU Day only		
ARDS Class	**Development Database: AUC, mean ± SD**	**Validation Database: AUC, mean ± SD**
Mild	*0.64 ± 0.06*	*0.55 ± 0.03*
Moderate	*0.61 ± 0.06*	*0.58 ± 0.05*
Severe	*0.62 ± 0.05*	*0.46 ± 0.05*
All	*0.63 ± 0.03*	*0.57 ± 0.02*
Event II: ICU mortality prediction in ARDS based on data of the 2nd ICU Day only		
ARDS Class	**Development Database: AUC, mean ± SD**	**Validation Database: AUC, mean ± SD**
Mild	*0.58 ± 0.10*	*0.59 ± 0.05*
Moderate	*0.61 ± 0.05*	*0.59 ± 0.03*
Severe	*0.53 ± 0.14*	*0.56 ± 0.05*
All	*0.63 ± 0.04*	*0.63 ± 0.03*
Event III: ICU mortality prediction in ARDS based on data of the 3rd ICU Day only		
ARDS Class	**Development Database: AUC, mean ± SD**	**Validation Database: AUC, mean ± SD**
Mild	*0.57 ± 0.09*	*0.55 ± 0.07*
Moderate	*0.62 ± 0.07*	*0.58 ± 0.08*
Severe	*0.50 ± 0.12*	*0.47 ± 0.06*
All	*0.63 ± 0.05*	*0.62 ± 0.05*

Table 5.4. Performance of RF models: P/FP_E results

Event I: ICU mortality prediction in ARDS based on data of the 1st ICU Day only		
ARDS Class	**Development Database: AUC, mean ± SD**	**Validation Database: AUC, mean ± SD**
Mild	*0.67 ± 0.05*	*0.57 ± 0.05*
Moderate	*0.58 ± 0.03*	*0.56 ± 0.06*
Severe	*0.65 ± 0.03*	*0.48 ± 0.04*
All	*0.64 ± 0.03*	*0.58 ± 0.01*
Event II: ICU mortality prediction in ARDS based on data of the 2nd ICU Day only		
ARDS Class	**Development Database: AUC, mean ± SD**	**Validation Database: AUC, mean ± SD**
Mild	*0.62 ± 0.09*	*0.61 ± 0.07*
Moderate	*0.59 ± 0.05*	*0.59 ± 0.05*
Severe	*0.65 ± 0.04*	*0.59 ± 0.07*
All	*0.64 ± 0.03*	*0.64 ± 0.03*
Event III: ICU mortality prediction in ARDS based on data of the 3rd ICU Day only		
ARDS Class	**Development Database: AUC, mean ± SD**	**Validation Database: AUC, mean ± SD**
Mild	*0.58 ± 0.13*	*0.56 ± 0.09*
Moderate	*0.56 ± 0.08*	*0.54 ± 0.08*
Severe	*0.66 ± 0.05*	*0.63 ± 0.06*
All	*0.64 ± 0.04*	*0.63 ± 0.05*

For the development database, the best ICU mortality predictive model for the mild class was obtained by event I and by P/FP_E with AUC = 0.67. When PaO_2/FiO_2 was used, the average AUC was 0.64. For the severe class, the best predictive model for ICU mortality was observed for event III and by P/FP_E with AUC = 0.66, outperforming PaO_2/FiO_2 whose AUC was 0.50. For this severe class, the AUC of the predictive models for ICU mortality by events I & II and by P/FP_E had the same value and was equal to 0.65. For the moderate class, the best predictive models for ICU mortality were observed for event III and by PaO_2/FiO_2 with AUC = 0.62, and for event II by P/FP_E with AUC = 0.59. For all ARDS patients, the best predictive model for ICU mortality was the one using P/FP_E with the same AUC value in all events (AUC = 0.64).

For the validation database, the best AUC values were 0.61 and 0.63 for P/FP_E in the mild and severe classes for events II and III, respectively; and 0.59 and 0.56 for PaO_2/FiO_2 in the

same classes for event II only. For the moderate class, the best predictive models for ICU mortality using both indices were observed for event II only and had the same AUC value of 0.59. For all ARDS patients, the best predictive model for ICU mortality was observed for event II and by P/FP_E with AUC = 0.64.

In general, the P/FP_E index had a better performance for predicting ICU mortality for mild and severe ARDS than the PaO_2/FiO_2 ratio. Whereas PaO_2/FiO_2 reached up to 0.47 AUC for severe class in the validation database, the use of P/FP_E reached 0.63 AUC. This represents an increment of +0.16 AUC, and quantifies the advantage of using the P/FP_E index as a predictor of ICU mortality in severe ARDS.

5.4 Discussion

In this large study, our major finding is that P/FP_E index is a more sensitive predictor of ICU mortality in severe, mild categories, and for all ARDS patients over time than PaO_2/FiO_2 ratio. Using ML, ICU mortality predictive models for PaO_2/FiO_2 ratio vs. P/FP_E index were generated in over 2,400 ARDS patients and externally validated in over 5,000 patients over time. Our findings reinforced our original study describing that P/FP_E is a more sensitive descriptor of severity of respiratory failure over time than PaO_2/FiO_2 [59].

In the development dataset, the RF model using the P/FP_E index in severe ARDS attained outstanding improvement in terms of AUC (16% improvement), when compared to the PaO_2/FiO_2 model. For the validation dataset, the RF model based on P/FP_E in the severe class also outperformed PaO_2/FiO_2 prediction and quantified the same improvement of AUC (i.e., 16% improvement) as in the development dataset. For both the development and validation datasets, best prediction of ICU outcome using P/FP_E index in severe ARDS was at *day-3*, a finding which is in accordance with previous studies [85,86].

Although we acknowledge that the best AUCs values for the P/FP_E index were not remarkably high, our findings should be interpreted within the following context. First, P/FP_E is a severity index and was not originally designed as a predictor of ICU mortality [59]. In other words, evaluating P/FP_E as a measure of ARDS severity by comparing its ability to predict overall ICU mortality may be problematic due to that this outcome may not be related to ARDS severity in several cases [76,78]. In this regard, the use of a minimum PEEP level of 5 cmH_2O recommended by the P/FP_E index might not be appropriate to allow reliable ICU mortality prediction [85]. Second, since P/FP_E index was internally derived from calculations using the PaO_2/FiO_2 ratio and applied PEEP (i.e., $(PaO_2/FiO_2)/PEEP$), it is not surprising that its ability for ICU mortality prediction may decrease in the external validation dataset [59]. Third, in general, ICU mortality as an outcome measure is not completely reflective of morality from ARDS since only a minority of ARDS patients die of severe lung injury and an inability to oxygenate and ventilate. Mortality in ARDS is primarily related to organ failures and the underlying etiology of ARDS such as sepsis [76,78,87]. Also, it must be acknowledged that the results of the RF model using the PaO_2/FiO_2 ratio in the moderate class were slightly better than those obtained by the P/FP_E index. However, further research will be necessary to investigate the wide heterogeneity within the moderate class [88], which includes patients with shunt fractions that may range from 20% to 60% and had the higher number of ARDS patients in both datasets [89]. Specifically, in future research we will investigate whether a P/FP_E cut-off of 30-mmHg/cmH_2O, may be a cut-off value in moderate ARDS for splitting this category into mild–moderate (P/FP_E between 30 and 40 mmHg/cmH_2O) and moderate–severe (P/FP_E between 21 and 29 mmHg/cmH_2O) groups. It is plausible that using the P/FP_E threshold of 30-mmHg/cmH_2O might achieve a more homogeneous distribution of ARDS patients across the severity subgroups and may identify two populations that differ in their anatomical and physiological characteristics.

Of note, the current tendency for publishing only overoptimistic ML results related to medicine that "improve upon the state-of-the-art" is an increasing serious problem. This problem leads to biased situations where algorithms and parameter sets are tuned and re-tuned almost indefinitely towards an as-high-as-possible performance [27,28].

The P/FP_E index outperformed PaO_2/FiO_2 ratio in all RF models for ICU mortality prediction in severe, mild classes, and all ARDS patients over time for both development and validation datasets. Villar et al. showed that the value of PaO_2/FiO_2 at 24 h after ARDS onset stratified patients with wide differences in mortality when the PaO_2/FiO_2 ratio was calculated under standardized ventilatory settings ($FiO_2 \geq 0.5$ with PEEP$\geq$10 cmH_2O) [90]. The P/FP_E was, however, not part of the Villar et al. study.

ARDS is a typical example of a very heterogeneous ICU syndrome, rather than a distinct disease. Such heterogeneity is challenging and translates into a wide range of severity, and partially explains the difficulty in investigating new therapies and contributes to the unprecedented high list of unsuccessful interventional trials in ARDS patients [35]. ARDS is a predominantly clinical diagnosis, but there have been difficulties in agreeing on a standardized, universal definition [91]. The PaO_2/FiO_2 ratio classifies the severity of ARDS based on the degree of the oxygenation deficit and within each of those categories, patients are assumed to be less heterogeneous [6]. However, recent observations by Pham et al. [36] showed that mild ARDS is underappreciated because it had a high mortality rate [35]. Since PaO_2/FiO_2 ratio does not account for PEEP in its calculation, it provides an incomplete picture of actual ARDS severity. Our study provides a solution for that dilemma by using the P/FP_E index, taking applied PEEP into account and creating three grades of severity based on intuitive classification thresholds that are different from the Berlin definition, and within each of these ARDS categories, the patients had a similar degree of lung severity [59].

Predicting ICU mortality is important for providing critical care physicians enough insights to make decisions and to allocate ICU resources. Therefore, predicting mortality within the first ICU days is a difficult task but a paramount endeavor [87]. Previous studies have reported the effectiveness of ML in predicting ICU mortality or other ICU events [59,60,82,87,92]. In this large study, a comparison of conventional ICU scores or traditional statistical regression-based techniques with RF-based algorithm, leads to important advantages for the RF algorithm: (i) the RF algorithm automatically learns the interaction and non-linear effects among the predictors from the data, and thus is more relevant for big data; (ii) after a 10-fold cross-validation, we obtained reliable performance AUCs in ML-standard as (mean ± SD) [27,28] for the RF models.

Our study has several strengths. First, to the best of our knowledge, this is the first study investigating ICU mortality prediction using the intuitive classification thresholds of P/FP_E index (different from the Berlin definition) based on ML-based models for each severity grade of ARDS over time, especially for more than two ICU days. Second, we have implemented and externally validated the RF models that can accurately predict ICU mortality for each severity grade of ARDS using two large ICU datasets (the single-center MIMIC-III and the eICU multicenter, respectively). Third, we included variables of arterial oxygenation and ventilator settings that were readily available in routine clinical practice to guarantee clinical relevance within a wide range of ARDS severity. This provides a readily available tool with bedside applicability and does not involve computation of complex equations with multiple variables into account [59,83-86]. Forth, we believe that our study is an important addition to the current ARDS heterogeneous picture. We attempt to bridge the gap between real world day to day application of MV with the multifactorial nature of mortality in ARDS. However, we acknowledge that our study has some limitations. First, this is a retrospective analysis of data whose results for the P/FP_E index privileges should be confirmed in further prospective studies.

Second, there was no standardized protocol for ventilator settings between the two ICU databases. Third, our analysis for the P/FP_E index is concerned with the progression and stratification of ARDS patients in three consecutive ICU time periods (first 24, 48, and 72 hours after the onset), and it is plausible that its prognostic ability would improve after 72 h of ARDS onset.

5.5 Conclusions

For both development and validation datasets, the RF models of severe, mild categories, and all ARDS patients achieved better performance using the P/FP_E index over time compared to those using the PaO_2/FiO_2 ratio. P/FP_E appears to be particularly helpful for prediction of ICU mortality in severe ARDS at *day-3* after onset.

CHAPTER 6

Concluding Remarks

This chapter presents the most significant contributions and main conclusions of this dissertation, highlighting their significance. Furthermore, the chapter also involves proposals for future work.

6.1 Conclusions

In this thesis, different methods for assessing ARDS have been presented. In particular, we have proposed a novel criteria and ML approach for ARDS severity classification (Chapter 3), we have developed efficient ML models for MV duration prediction (Chapter 4), and provided mortality prediction models (Chapter 5) based on the P/FP_E that outperformed previous models defined under the Berlin definition for severe and mild ARDS patients.

Usually, arterial oxygenation in patients with ARDS improves substantially by increasing the level of PEEP. In **Chapter 3**, we have proposed a novel formula [$PaO_2/(FiO_2$xPEEP) or P/FP_E] for PEEP≥5 and corresponding cut-off values to address Berlin's definition gap for ARDS severity by using ML approaches. We examined P/FP_E values delimiting the boundaries of mild, moderate, and severe ARDS. We applied ML to predict ARDS severity after onset over time by comparing current Berlin PaO_2/FiO_2 ratio with P/FP_E index under three different scenarios. Three robust

classification algorithms were implemented using Python 3.7 (and the classifiers RF, XGBoost and lightGBM) for predicting ARDS severity over time. P/FP_E index outperformed PaO_2/FiO_2 ratio in all ARDS severity prediction models after onset over time (MIMIC-III: AUC 0.711-0.788 and CORR 0.376-0.566; eICU: AUC 0.734-0.873 and CORR 0.511-0.745). We found that the novel P/FP_E index to assess ARDS severity after onset over time is markedly better than current PaO_2/FiO_2 ratio.

Clinically, applying the proposed new criteria for ARDS severity enables critical care physicians to assess lung severity by involving PEEP information. Moreover, being able to better adjust the severity profiles of ARDS patients will potentially improve the selection of more adequate therapeutic regimens for each ARDS category, which could contribute to reduce ARDS mortality. From a ML perspective, P/FP_E outperformed PaO_2/FiO_2 in all ML models predicting ARDS severity after onset over time in all scenarios either in MIMIC-III or eICU. Accordingly, this study can serve as an example of how ML is a worth-considering technology to gain new insights in the development of ARDS predictive models which could contribute to improve ICU resource allocation and mortality reduction.

Most ARDS patients require MV. Few studies have investigated the prediction of MV duration over time. In **Chapter 4**, we aimed at characterizing the best early scenario during the first two days in ICU to predict MV duration after ARDS onset using supervised machine learning approaches. LightGBM was the best model in predicting MV duration after ARDS onset in MIMIC-III with a RMSE of 6.10–6.41 days, and it was externally validated in eICU with RMSE of 5.87–6.08 days. The best early prediction model was obtained with data captured in the 2nd day. With our work, we have proved that supervised ML models can accurately and early predict MV duration in ARDS after onset over time across ICUs and it opens the possibility of reaching

new better MV duration prediction models based on ML. Supervised ML models might have important implications for optimizing ICU resource utilization and high acute cost reduction of MV.

Chapter 5 was designed to validate P/FP_E as a predictor of ICU mortality in two large databases using intuitive classification thresholds based on ML. In the development dataset, best performance was on ICU day-3 in severe ARDS (AUC=0.66), on ICU day-1 in mild ARDS (AUC=0.67), and in all ARDS patients (AUC=0.64) in all the three ICU days. In the validation dataset, the best AUC=0.63 was on ICU day-3 in severe ARDS, the best AUC=0.61 was on ICU day-2 in mild ARDS, and the best AUC=0.64 was on ICU day-2 in all ARDS patients. We found that P/FP_E index is a more sensitive predictor of ICU mortality over time than the PaO_2/FiO_2 ratio in all ARDS categories.

6.2 Future work

The work presented in this thesis makes a substantial contribution to the assessment of ARDS severity by using ML. We believe this is an addition to the story of ARDS research. An important direction of future work can be suggested as the application of deep learning techniques for building robust ARDS assessment models.

On the other hand, evaluation of ML solutions is a multifaceted process that needs the expertise of data scientists, clinician experts and implementation scientists. Currently, most studies using evaluation of these solutions remain focused on internal validation, with relatively few studies examining clinical outcomes and system implementation. This imbalance has contributed to what is referred to as the "AI chasm," representing the gap between the development and validation of ML algorithms and their eventual application in the clinical practice. Thus, additional clinical

outcomes and implementation research is necessary to fully realize the potential of ML in supporting decision making in healthcare [27,28].

References

1. Bellani G, Laffey JG, Pham T, Fan E, Brochard L, Esteban A, et al. Epidemiology, patterns of care, and mortality for patients with acute respiratory distress syndrome in Intensive Care Units in 50 countries. JAMA. 2016;315(8):788–800. https://doi.org/10.1001/jama.2016.0291

2. Thompson BT, Chambers RC, Liu KD. Acute Respiratory Distress Syndrome. N Engl J Med. 2017;377(19):1904–05. https://doi.org/10.1056/nejmc1711824

3. Calfee CS, Delucchi K, Parsons PE, Thompson BT, Ware LB, Matthay MA. Subphenotypes in acute respiratory distress syndrome: latent class analysis of data from two randomised controlled trials. Lancet Respir Med. 2014;2(8):611-20. https://doi.org/10.1016/S2213-2600(14)70097-9

4. Sinha P, Delucchi KL, Thompson BT, McAuley DF, Matthay MA, Calfee CS. Latent class analysis of ARDS subphenotypes: a secondary analysis of the statins for acutely injured lungs from sepsis (SAILS) study. Intensive Care Med. 2018;44(11):1859-69. https://doi.org/10.1007/s00134-018-5378-3

5. Del Sorbo L, Ranieri VM, Ferguson ND. The Berlin definition met our needs: yes. Intensive Care Med. 2016;42(5):643–47. https://doi.org/10.1007/s00134-016-4286-7

6. ARDS Definition Task Force, Ranieri VM, Rubenfeld GD, Thompson BT, Ferguson ND, Caldwell E, Fan E, et al. Acute respiratory distress syndrome: the Berlin definition. JAMA. 2012;307(23):2526-33. https://doi.org/10.1001/jama.2012.5669

7. Ferguson ND, Fan E, Camporota L, Antonelli M, Anzueto A, Beale R, et al. The Berlin definition of ARDS: an expanded rationale, justification, and supplementary material. Intensive Care Med. 2012;38(10):1573–82. https://doi.org/10.1007/s00134-012-2682-1

8. Villar J, Perez-Mendez L, Kacmarek RM. The Berlin definition met our needs: no. Intensive Care Med. 2016;42(5):648–50. https://doi.org/10.1007/s00134-016-4242-6

9. Bernard GR, Artigas A, Brigham KL, Carlet J, Falke K, Hudson L, et al. Report of the American-European consensus conference on ARDS: definitions, mechanisms, relevant outcomes and clinical trial coordination. Intensive Care Med. 1994;20(3):225–32. https://doi.org/10.1007/bf01704707

10. Costa EL, Amato MB. The new definition for acute lung injury and acute respiratory distress syndrome: is there room for improvement? Curr Opin Crit Care. 2013;19(1):16–23. https://doi.org/10.1097/MCC.0b013e32835c50b1

11. Thompson BT, Matthay MA. The Berlin definition of ARDS versus pathological evidence of diffuse alveolar damage. Am J Respir Crit Care Med. 2013;187(7):675-77. https://doi.org/10.1164/rccm.201302-0385ed

12. Villar J, Blanco J, del Campo R, Andaluz-Ojeda D, Díaz-Domínguez FJ, Muriel A, et al. Assessment of PaO_2/FiO_2 for stratification of patients with moderate and severe acute respiratory distress syndrome. BMJ Open. 2015;5(3):e006812. https://doi.org/10.1136/bmjopen-2014-006812

13. Huber W, Findeisen M, Lahmer T, Herner A, Rasch S, Mayr U, et al. Prediction of outcome in patients with ARDS: A prospective cohort study comparing ARDS-definitions and other ARDS-associated parameters, ratios and scores at intubation and over time. PLoS ONE. 2020;15(5)e0232720. https://doi.org/10.1371/journal.pone.0232720

14. Moss M, Huang DT, Brower RG, Ferguson ND, Ginde AA, Gong MN, et al. Early Neuromuscular Blockade in the Acute Respiratory Distress Syndrome. N Engl J Med. 2019;380(21):1997-2008. https://doi.org/10.1056/NEJMoa1901686

15. Slutsky AS, Villar J. Early Paralytic Agents for ARDS? Yes, No, and Sometimes. N Engl J Med. 2019;380(21):2061-63. https://doi.org/10.1056/NEJMe1905627

16. Phillips CR. The Berlin definition: real change or the emperor's new clothes? Crit Care. 2013; 17(4):174. https://doi.org/10.1186/cc12761

17. Allardet-Servent J, Forel JM, Roch A, Guervilly C, Chiche L, Castanier M, et al. FiO_2 and acute respiratory distress syndrome definition during lung protective ventilation. Crit. Care Med. 2009; 37(1):202–07, e4-6. https://doi.org/10.1097/CCM.0b013e31819261db

18. Balzer F, Menk M, Ziegler J, Pille C, Wernecke KD, Spies C, et al. Predictors of survival in critically ill patients with acute respiratory distress syndrome (ARDS): an observational study. BMC Anesthesiol. 2016; 16(1):108. https://doi.org/10.1186/s12871-016-0272-4

19. Dai Q, Wang S, Liu R, Wang H, Zheng J, Yu K. Risk factors for outcomes of acute respiratory distress syndrome patients: a retrospective study. J Thorac Dis. 2019; 11(3):673–85. https://doi.org/10.21037/jtd.2019.02.84

20. Kamo T, Tasaka S, Suzuki T, Asakura T, Suzuki S, Yagi K, et al. Prognostic values of the Berlin definition criteria, blood lactate level, and fibroproliferative changes on high-resolution computed tomography in ARDS patients. BMC Pulm Med. 2019; 19(1):37. https://doi.org/10.1186/s12890-019-0803-0

21. Lai CC, Sung MI, Liu HH, Chen CM, Chiang SR, Liu WL, et al. The ratio of partial pressure arterial oxygen and fraction of inspired oxygen 1 day after acute respiratory distress syndrome onset can predict the outcomes of involving patients. Medicine (Baltimore). 2016;95(14):e3333. https://doi.org/10.1097/MD.0000000000003333

22. Rubenfeld GD, Caldwell E, Peabody E, Weaver J, Martin DP, Neff M, et al. Incidence and outcomes of acute lung injury. N Engl J Med. 2005;353(16):1685-93. https://doi.org/10.1056/NEJMoa050333

23. Slutsky AS, Villar J, Pesenti A. Happy 50th birthday ARDS! Intensive Care Med. 2016;42(5):637-39. https://doi.org/10.1007/s00134-016-4284-9

24. Bice T, Carson SS. Acute Respiratory Distress Syndrome: Cost (Early and Long-Term). Seminars in Respiratory and Critical Care Medicine. 2019;40(1):137-144. https://doi.org/10.1055/s-0039-1685463

25. Dasta JF, McLaughlin TP, Mody SH, Piech CT. Daily cost of an intensive care unit day: the contribution of mechanical ventilation. *Crit. Care Med*. 2005;33(6):1266-71. https://doi.org/10.1097/01.ccm.0000164543.14619.00

26. Marti J, Hall P, Hamilton P, Lamb S, McCabe C, Lall R, et al. One-year resource utilisation, costs and quality of life in patients with acute respiratory distress syndrome (ARDS): secondary analysis of a randomised controlled trial. J Intensive Care. 2016;4:56. https://doi.org/10.1186/s40560-016-0178-8

27. Tohka J, van Gils M. Evaluation of machine learning algorithms for health and wellness applications: A tutorial. Comput Biol Med. 2021;132:104324. http://dx.doi.org/10.1016/j.compbiomed.2021.104324

28. Antoniou T, Mamdani M. Evaluation of machine learning solutions in medicine. CMAJ. 2021;193(36):E1425-29. https://doi.org//10.1503/cmaj.210036

29. Boulesteix AL, Janitza S, Kruppa J, Konig IR. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Int. Rev. Data Min. and Knowl. Discov. 2012;2(6)493-507. http://dx.doi.org/10.1002/widm.1072

30. Chen T, and Carlos G. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd *ACM SIGKDD* International Conference on Knowledge Discovery and Data Mining *(KDD '16).* Association for Computing Machinery, New York, NY, USA. 2016;785–94. https://doi.org/10.1145/2939672.2939785

31. Dehua W, Yang Z, Yi Z. LightGBM: An effective miRNA classification method in breast cancer patients. In Proceedings of the 2017 International Conference on Computational Biology and Bioinformatics (*ICCBB* 2017). ACM, New York, NY, USA, 2017;7-11. https://doi.org//10.1145/3155077.3155079

32. Physionet.org, 'MIMIC-III Critical Care Database'. [Online]. Available: https://mimic.physionet.org/about/mimic/. (Accessed 2 July, 2020).

33. Physionet.org, 'eICU Collaborative Research Database'. [Online]. Available: https://eicu-crd.mit.edu/about/eicu/ (Accessed 19 October, 2020).

34. Hernu R, Wallet F, Thiolliére F, Martin O, Richard JC, Schmitt Z, et al. An attempt to validate the modification of the American-European consensus definition of acute lung injury/acute respiratory distress syndrome by the Berlin definition in a university hospital. Intensive Care Med. 2013;39(12):2161–70. https://doi.org/10.1007/s00134-013-3122-6

35. Pirracchio R, Gropper MA. Heterogeneity in Intensive Care: low severity does not mean low risk! Anesthesiology. 2019;130(2):190-91. https://doi.org/10.1097/ALN.0000000000002537

36. Pham T, Serpa Neto A, Pelosi P, Laffey JG, De Haro C, Lorente JA, et al. Outcomes of patients presenting with mild acute respiratory distress syndrome: insights from the LUNG SAFE

study. Anesthesiology. 2019;130(2):263-83. https://doi.org/10.1097/ALN.0000000000002508

37. Maiolo G, Collino F, Vasques F, Rapetti F, Tonetti T, Romitti F, et al. Reclassifying acute respiratory distress syndrome. Am J Respir Crit Care Med. 2018;197(12):1586–95. https://doi.org/10.1164/rccm.201709-1804OC
38. Bourenne J, Carvelli J, Papazian L. Evolving definition of acute respiratory distress syndrome. J Thorac Dis. 2019;11(Suppl 3):S390-S393. https://doi.org/10.21037/jtd.2018.12.24
39. Villar J, Fernández RL, Ambrós A, Parra L, Blanco J, Domínguez-Berrot AM, et al. A clinical classification of the acute respiratory distress syndrome for predicting outcome and guiding medical therapy. Crit Care Med. 2015;43(2):346-53. https://doi.org/10.1097/CCM.0000000000000703
40. Ferring M, Vincent JL. Is outcome from ARDS related to the severity of respiratory failure? Eur Respir J. 1997;10(6):1297–300. https://doi.org/10.1183/09031936.97.10061297
41. Villar J, Pérez-Méndez L, López J, Belda J, Blanco J, Saralegui I, et al. An early PEEP/FIO2 trial identifies different degrees of lung injury in patients with acute respiratory distress syndrome. Am J Respir Crit Care Med. 2007;176(8):795-804. https://doi.org/10.1164/rccm.200610-1534OC
42. Jia X, Malhotra A, Saeed M, Mark RG, Talmor D. Risk factors for ARDS in patients receiving mechanical ventilation for >48 h. Chest. 2008;133(4):853-61. https://doi.org/10.1378/chest.07-1121
43. Mahmoud O. Mechanical power is associated with increased mortality and worsened oxygenation in ARDS. 2020;Chest,ISSN: 0012-3692,158(4), A679.

44. Monchi M, Bellenfant F, Cariou A, Joly LM, Thebert D, Laurent I, et al. Early predictive factors of survival in the acute respiratory distress syndrome. A multivariate analysis. Am J Respir Crit Care Med. 1998;158(4):1076-81. https://doi.org/10.1164/ajrccm.158.4.9802009

45. Pintado MC, de Pablo R, Trascasa M, Milicua JM, Rogero S, Daguerre M, et al. Individualized PEEP setting in subjects with ARDS: a randomized controlled pilot study. Respir Care. 2013;58(9):1416-23. https://doi.org/10.4187/respcare.02068

46. Le S, Pellegrini E, Green-Saxena A, Summers C, Hoffman J, Calvert J, et al. Supervised machine learning for the early prediction of acute respiratory distress syndrome (ARDS). J Crit Care. 2020;60:96-102. https://doi.org/10.1016/j.jcrc.2020.07.019

47. Flaatten H, Gjerde S, Guttormsen AB, Haugen O, Hoivik T, Onarheim H, et al. Outcome after acute respiratory failure is more dependent on dysfunction in other vital organs than on the severity of the respiratory failure. Crit Care. 2003;7(4):R72. https://doi.org/10.1186/cc2331

48. Austin PC, Tu JV, Ho JE, Levy D, Lee DS. Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. J Clin Epidemiol. 2013;66(4):398-407. https://doi.org/10.1016/j.jclinepi.2012.11.008

49. Hosmer W, Lemeshow JR. Applied Logistic Regression. John Wiley & Sons, New York, 2004.

50. Mukaka MM. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. Malawi Med J. 2012;24(3):69-71.

51. Ding XF, Li JB, Liang HY, Wang ZY, Jiao TT, Liu Z, et al. Predictive model for acute respiratory distress syndrome events in ICU patients in China using machine learning algorithms: a secondary analysis of a cohort study. J Transl Med. 2019;17(1):326. https://doi.org/10.1186/s12967-019-2075-0

52. Zeiberg D, Prahlad T, Nallamothu BK, Iwashyna TJ, Wiens J, Sjoding MW. Machine learning for patient risk stratification for acute respiratory distress syndrome. PLoS One. 2019;14(3):e0214465. https://doi.org/10.1371/journal.pone.0214465

53. Rush B, Stone DJ, Celi LA. From Big Data to Artificial Intelligence: Harnessing Data Routinely Collected in the Process of Care. Crit Care Med. 2018;46(2):345-46. https://doi.org/10.1097/CCM.0000000000002892

54. Figueroa-Casas JB, Connery SM, Montoya R, Dwivedi AK, Lee S. Accuracy of early prediction of duration of mechanical ventilation by intensivists. Ann Am Thorac Soc. 2014;11(2):182-5. https://doi.org/10.1513/AnnalsATS.201307-222OC

55. Figueroa-Casas JB, Dwivedi AK, Connery SM, Quansah R, Ellerbrook L, Galvis J. Predictive models of prolonged mechanical ventilation yield moderate accuracy. J. Crit. Care. 2015;30(3):502-5. https://doi.org/10.1016/j.jcrc.2015.01.020

56. Cherifa M, Pirracchio R. What every intensivist should know about Big Data and targeted machine learning in the intensive care unit. Rev Bras Ter Intensiva. 2019;31(4):444-46. https://doi.org/10.5935/0103-507X.20190069

57. Gutierrez G. Artificial Intelligence in the Intensive Care Unit. Crit Care. 2020;24(1):101. http://doi.org/10.1186/s13054-020-2785-y

58. Greco M, Caruso PF, Cecconi M. Artificial Intelligence in the Intensive Care Unit. Semin Respir Crit Care Med. 2021;42(1):2-9. https://doi.org/10.1055/s-0040-1719037

59. Sayed M, Riaño D, Villar J. Novel criteria to classify ARDS severity using a machine learning approach. Crit Care. 2021; 25(1):150. https://doi.org/10.1186/s13054-021-03566-w

60. Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der Laan MJ. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a

population-based study. Lancet Respir Med. 2015;3(1):42-52. https://doi.org/10.1016/S2213-2600(14)70239-5

61. Troché G, Moine P. Is the duration of mechanical ventilation predictable? Chest. 1997;112(3):745-51. https://doi.org/10.1378/chest.112.3.745

62. Parreco J, Hidalgo A, Parks JJ, Kozol R, Rattan R. Using artificial intelligence to predict prolonged mechanical ventilation and tracheostomy placement. J Surg Res. 2018;228:179-87. https://doi.org/10.1016/j.jss.2018.03.028

63. Sayed M, Riaño D. Modelling ICU Patients to Improve Care Requirements and Outcome Prediction of Acute Respiratory Distress Syndrome: A Supervised Learning Approach. In: Marcos M. et al., eds. Artificial Intelligence in Medicine: Knowledge Representation and Transparent and Explainable Systems. *KR4HC-ProHealth 2019/TEAAM 2019, LNAI*. 2019;11979:39-49. Springer, Cham. https:// doi.org/10.1007/978-3-030-37446-4_4

64. Villar J, Ambrós A, Soler JA, et al. Age, PaO_2/FiO_2, and Plateau Pressure Score: A Proposal for a Simple Outcome Score in Patients With the Acute Respiratory Distress Syndrome. *Crit Care Med*. 2016;44(7):1361-69. https://doi.org/10.1097/CCM.0000000000001653

65. Gong MN, Schenk L, Gajic O, Mirhaji P, Sloan J, Dong Y, et al. Early intervention of patients at risk for acute respiratory failure and prolonged mechanical ventilation with a checklist aimed at the prevention of organ failure: protocol for a pragmatic stepped-wedged cluster trial of PROOFCheck. BMJ Open. 2016;6(6):e011347. https://doi.org/10.1136/bmjopen-2016-011347

66. Hagan R, Gillan CJ, Spence I, McAuley D, Shyamsundar M. Comparing regression and neural network techniques for personalized predictive analytics to promote lung protective ventilation

in Intensive Care Units. Comput Biol Med. 2020;126:104030. https://doi.org/10.1016/j.compbiomed.2020.104030

67. Marco L. Intensive care resource allocation: When difficult choices have to be made. BJMP. 2013;6(4):a633.

68. Seneff MG, Zimmerman JE, Knaus WA, Wagner DP, Draper EA. Predicting the duration of mechanical ventilation. The importance of disease and patient characteristics. Chest. 1996;110(2):469-79. https://doi.org/10.1378/chest.110.2.469

69. Marshall Y, Hayley B, Chao-Ping Wu, Michelle N. Increased Economic Costs Associated with Acute Respiratory Distress Syndrome in Mechanically Ventilated Patients in the Intensive Care Unit. Am. J. Respir. Crit. Care Med. 2017;195:A7579.

70. Sheikhalishahi S, Balaraman V, Osmani V. Benchmarking machine learning models on multi-centre eICU critical care dataset. PLoS One. 2020;15(7):e0235424. https://doi.org/10.1371/journal.pone.0235424

71. Laffey JG, Bellani G, Pham T, Fan E, Madotto F, Bajwa EK, et al. Potentially modifiable factors contributing to outcome from acute respiratory distress syndrome: the LUNG SAFE study. Intensive Care Med. 2018;44(1):157-65. https://doi.org/10.1007/s00134-017-4981-z

72. Rosenberg AL, Dechert RE, Park PK, Bartlett RH; NIH NHLBI ARDS Network. Review of a large clinical series: association of cumulative fluid balance on outcome in acute lung injury: a retrospective review of the ARDSnet tidal volume study cohort. J Intensive Care Med. 2009;24(1):35-46. https://doi.org/10.1177/0885066608329850

73. Gong MN, Thompson BT, Williams P, Pothier L, Boyce PD, Christiani DC. Clinical predictors of and mortality in acute respiratory distress syndrome: potential role of red cell transfusion. Crit Care Med. 2005;33(6):1191-8. https://doi.org/10.1097/01.ccm.0000165566.82925.14

74. DesPrez K, McNeil JB, Wang C, Bastarache JA, Shaver CM, Ware LB. Oxygenation Saturation Index Predicts Clinical Outcomes in ARDS. Chest. 2017;152(6):1151-58. https://doi.org/10.1016/j.chest.2017.08.002

75. Sinha P, Calfee CS, Beitler JR, Soni N, Ho K, Matthay MA, et al. Physiologic Analysis and Clinical Performance of the Ventilatory Ratio in Acute Respiratory Distress Syndrome. Am J Respir Crit Care Med. 2019;199(3):333-341. https://doi.org/10.1164/rccm.201804-0692OC

76. Sedhai YR, Yuan M, Ketcham SW, Co I, Claar DD, McSparron JI, et al. Validating Measures of Disease Severity in Acute Respiratory Distress Syndrome. Ann Am Thorac Soc. 2021;18(7):1211-1218. https://doi.org/10.1513/AnnalsATS.202007-772OC

77. Villar J, González-Martín JM, Ambrós A, Mosteiro F, Martínez D, Fernández L, et al. Stratification for Identification of Prognostic Categories In the Acute RESpiratory Distress Syndrome (SPIRES) Score. Crit Care Med. 2021;49(10):e920-e930. https://doi.org/10.1097/CCM.0000000000005142

78. Ketcham SW, Sedhai YR, Miller HC, Bolig TC, Ludwig A, Co I, et al. Causes and characteristics of death in patients with acute hypoxemic respiratory failure and acute respiratory distress syndrome: a retrospective cohort study. Crit Care. 2020;24(1):391. https://doi.org/10.1186/s13054-020-03108-w

79. Kacmarek RM, Berra L. Prediction of ARDS outcome: what tool should I use? Lancet Respir Med. 2018;6(4):253-54. https://doi.org/10.1016/S2213-2600(18)30098-5

80. Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, et al. Prediction of Sepsis in the Intensive Care Unit with Minimal Electronic Health Record Data: A Machine Learning Approach. JMIR Med Inform. 2016 30;4(3):e28. https://doi.org/10.2196/medinform.5909

81. Yoon JH, Pinsky MR, Clermont G. Artificial Intelligence in Critical Care Medicine. Crit Care. 2022;26(1):75. https://doi.org/10.1186/s13054-022-03915-3

82. Sayed M, Riaño D, Villar J. Predicting Duration of Mechanical Ventilation in Acute Respiratory Distress Syndrome Using Supervised Machine Learning. J Clin Med. 2021;10(17):3824. https://doi.org/10.3390/jcm10173824

83. Mamandipoor B, Frutos-Vivar F, Peñuelas O, Rezar R, Raymondos K, Muriel A, et al. Machine learning predicts mortality based on analysis of ventilation parameters of critically ill patients: multi-centre validation. BMC Med Inform Decis Mak. 2021;21(1):152. https://doi.org/10.1186/s12911-021-01506-w

84. Wernly B, Mamandipoor B, Baldia P, Jung C, Osmani V. Machine learning predicts mortality in septic patients using only routinely available ABG variables: a multi-centre evaluation. Int J Med Inform. 2021;145:104312. https://doi.org/10.1016/j.ijmedinf.2020.104312

85. Balzer F, Menk M, Ziegler J, Pille C, Wernecke KD, Spies C, et al. Predictors of survival in critically ill patients with acute respiratory distress syndrome (ARDS): an observational study. BMC Anesthesiol. 2016 Nov 8;16(1):108. http://dx.doi.org/10.1186/s12871-016-0272-4

86. Chiu LC, Lin SW, Liu PH, Chuang LP, Chang CH, Hung CY, et al. Reclassifying severity after 48 hours could better predict mortality in acute respiratory distress syndrome. Ther Adv Respir Dis. 2020;14:1753466620936877. http://dx.doi.org/10.1177/1753466620936877

87. Wong AI, Cheung PC, Kamaleswaran R, Martin GS, Holder AL. Machine Learning Methods to Predict Acute Respiratory Failure and Acute Respiratory Distress Syndrome. Front Big Data. 2020;3:579774. http://dx.doi.org/10.3389/fdata.2020.579774

88. Cuadrado D, Riaño D, Gómez J, Rodríguez A, Bodí M. Methods and measures to quantify ICU patient heterogeneity. J Biomed Inform. 2021;117:103768. http://dx.doi.org/10.1016/j.jbi.2021.103768

89. Maiolo G, Collino F, Vasques F, Rapetti F, Tonetti T, Romitti F, et al. Reclassifying Acute Respiratory Distress Syndrome. Am J Respir Crit Care Med. 2018;197(12):1586-95. http://dx.doi.org/10.1164/rccm.201709-1804OC

90. Villar J, Pérez-Méndez L, Blanco J, Añón JM, Blanch L, Belda J, et al. A universal definition of ARDS: the PaO2/FiO2 ratio under a standard ventilatory setting--a prospective, multicenter validation study. Intensive Care Med. 2013;39(4):583-92. https://doi.org/10.1007/s00134-012-2803-x

91. McCormack V, Tolhurst-Cleaver S. Acute respiratory distress syndrome. BJA Education. 2017;17(5):161-65. https://doi.org/10.1093/bjaed/mkx002

92. Marafino BJ, Park M, Davies JM, Thombley R, Luft HS, Sing DC, et al. Validation of Prediction Models for Critical Care Outcomes Using Natural Language Processing of Electronic Health Record Data. JAMA Netw Open. 2018;1(8):e185097. https://doi.org/10.1001/jamanetworkopen.2018.5097

Appendix A

Table A1. Input variables and their descriptive statistics in MIMIC-III at 24-h according to PaO_2/FiO_2.

	Mild	**Moderate**	**Severe**	**All**
A. ARDS Patients	***669 (24.43%)***	***1,263(46.13%)***	***806 (29.44%)***	***2,738 (100%)***
*B. Descriptive feature– **means and 95% CI***				
Age	63.72 [62.52, 64.91]	62.70 [61.82, 63.58]	60.14 [58.97, 61.30]	62.19 [61.59, 62.80]
PEEP	5.87 [5.71, 6.02]	6.95 [6.79, 7.12]	9.57 [9.26, 9.87]	7.46 [7.32, 7.59]
Heart Rate_Mean	89 [87, 90]	91 [90, 92]	94 [93, 95]	91 [91, 92]
Respiratory Rate_Mean	19 [18, 19]	20 [19, 20]	22 [22, 23]	20 [20, 21]
Heart Rate_Max	110 [109, 112]	113 [112, 114]	117 [115, 118]	114 [113, 114]
Heart Rate_Min	73, [71, 74]	74 [73, 75]	77 [76, 78]	75 [74, 75]
Respiratory Rate_Max	27 [27, 28]	29 [29, 30]	32 [31, 32]	29 [29, 30]
Respiratory Rate_Min	12 [12, 13]	13 [12, 13]	14 [13, 14]	13 [12, 13]
SpO_2_Mean	98 [98, 99]	97 [97, 98]	96 [95, 96]	97 [97, 98]
SpO_2_Max	100 [99, 100]	100 [99, 100]	100 [99, 100]	100 [99, 100]
SpO_2_Min	92 [91, 93]	90 [89, 90]	86 [85, 87]	89 [89, 90]

Table A2. Input variables and their descriptive statistics in MIMIC-III at 48-h according to PaO_2/FiO_2.

	Mild	**Moderate**	**Severe**	**All**
A. ARDS Patients	***512 (33.71%)***	***778 (51.22%)***	***229 (15.08%)***	***1,519 (100%)***
*B. Descriptive feature– **means and 95% CI***				
Age	62.29 [60.89, 63.69]	60.45 [59.28, 61.61]	58.11 [55.92, 60.29]	60.72 [59.89, 61.55]
PEEP	7.32 [7.02, 7.62]	9.33 [8.94, 9.72]	11.91 [10.89, 12.94]	9.04 [8.76, 9.32]
Heart Rate_Mean	91 [89, 92]	93 [92, 94]	97 [95, 99]	93 [92, 94]
Respiratory Rate_Mean	20 [20, 21]	21 [20, 21]	23 [22, 23]	21 [20, 21]
Heart Rate_Max	113 [111, 115]	115 [114, 117]	121 [118, 124]	116 [114, 117]
Heart Rate_Min	74 [73, 76]	76 [75, 78]	78 [76, 80]	76 [75, 77]
Respiratory Rate_Max	29 [28, 30]	30 [29, 31]	33 [32, 34]	30 [29, 31]
Respiratory Rate_Min	13 [13, 14]	13 [13, 14]	14 [13, 14]	13 [13, 14]
SpO_2_Mean	98 [97, 98]	97 [96, 97]	96 [95, 96]	97 [96, 97]
SpO_2_Max	100 [99, 100]	100 [99, 100]	100 [99, 100]	100 [99, 100]
SpO_2_Min	90 [89, 91]	88 [87, 89]	85 [83, 86]	88 [88, 89]

Table A3. Other clinical outcomes and their descriptive statistics in MIMIC-III at 24-h, 48-h, and 72-h.

Database	ICU Day; (*N*)	Criteria	ARDS Class	ICU Mortality Rate (%) *Mean and 95% CI*	Duration of MV(hours) *Mean and 95% CI*
Mimic III	Day 1 (*669*)	*Berlin definition*	*Mild*	*24.07 [20.8, 27.3]*	*184.18 [173.77, 194.59]*
	Day 2 (*512*)			*23.44,[19.76, 27.12]*	*192.58 [180.27, 204.90]*
	Day 3 (*506*)			*23.72 [19.99, 27.44]*	*192.07 [180.92, 203.23]*
	Day 1 (*1,263*)		*Moderate*	*25.73 [23.3, 28.1]*	*194.32 [186.37, 202.27]*
	Day 2 (*778*)			*27.51 [24.36, 30.65]*	*212.81 [201.97, 223.64]*
	Day 3 (*678*)			*30.09 [26.63, 33.55]*	*222.54 [210.67, 234.41]*
	Day 1 (*806*)		*Severe*	*30.52 [27.3, 33.7]*	*217.73 [206.92, 228.54]*
	Day 2 (*229*)			*37.99 [31.66, 44.33]*	*245.44 [224.27, 266.62]*
	Day 3 (*157*)			*40.13 [32.38, 47.88]*	*245.77 [216.74, 274.79]*
	Day 1 (*554*)	*New severity criteria*	*Mild*	*25.5 [22, 29]*	*183.51 [171.95, 195.07]*
	Day 2 (*310*)			*26.5 [22, 31]*	*177.17 [163.07, 191.28]*
	Day 3 (*287*)			*25.1 [20, 30]*	*182.83 [168.10, 197.55]*
	Day 1 (*930*)		*Moderate*	*26.8 [24, 30]*	*185.27 [176.91, 193.63]*
	Day 2 (*482*)			*26.3 [22, 30]*	*187.23 [175.88, 198.58]*
	Day 3 (*453*)			*28.7 [25, 33]*	*190.26 [178.40, 202.12]*
	Day 1 (*1,254*)		*Severe*	*27.3 [25, 30]*	*215.45 [206.60, 224.29]*
	Day 2 (*727*)			*29.2 [26, 32]*	*240.99 [228.54, 253.45]*
	Day 3 (*601*)			*30.8 [27, 34]*	*246.25 [232.55, 259.95]*

Table A4. Input variables and their descriptive statistics in eICU at 24-h according to PaO_2/FiO_2.

	Mild	Moderate	Severe	All
A. ARDS Patients	***1,549 (30.06%)***	***2,361 (45.82%)***	***1,243 (24.12%)***	***5,153 (100%)***
*B. Descriptive feature– **means and 95% CI***				
Age	64.04 [63.28, 64.80]	64.14 [63.56, 64.73]	61.11 [60.27, 61.95]	63.38 [62.97, 63.79]
PEEP	5.72 [5.60, 5.83]	6.25 [6.14, 6.35]	8.55 [8.36, 8.75]	6.64 [6.56, 6.73]
FiO2	0.51 [0.50, 0.52]	0.59 [0.59, 0.61]	0.85 [0.84, 0.86]	0.63 [0.63, 0.64]
PaO2	128.05 [125.39, 130.72]	98.84 [97.54, 100.14]	84.19 [82.79, 85.59]	104.09 [102.94, 105.24]
PaCO2	40.74 [40.19, 41.29]	44.05 [43.52, 44.57]	45.99 [45.35, 46.64]	43.52 [43.19, 43.86]

Table A5. Input variables and their descriptive statistics in eICU at 48-h according to PaO_2/FiO_2.

	Mild	Moderate	Severe	All
A. ARDS Patients	***1,098 (36.83%)***	***1,454 (48.78%)***	***429 (14.39%)***	***2,981 (100%)***
*B. Descriptive feature– **means and 95% CI***				
Age	64.73 [63.85, 65.61]	63.29 [62.54, 64.06]	59.97 [58.67, 61.28]	63.35 [62.82, 63.88]
PEEP	6.11 [5.95, 6.27]	7.05 [6.91, 7.19]	9.93 [9.55, 10.31]	7.12 [7.00, 7.23]
FiO_2	0.42 [0.41, 0.43]	0.53 [0.52, 0.53]	0.83 [0.81, 0.85]	0.53 [0.52, 0.54]
PaO_2	102.31 [100.52, 104.09]	83.33 [82.23, 84.43]	74.73 [72.82, 76.63]	89.08 [88.12, 90.05]
$PaCO_2$	39.46 [38.91, 40.01]	41.89 [41.34, 42.44]	44.02 [42.94, 45.09]	41.30 [40.93, 41.68]

Table A6. Other clinical outcomes and their descriptive statistics in eICU at 24-h, 48-h, and 72-h.

Database	ICU Day; (*N*)	Criteria	ARDS Class	ICU Mortality Rate (%) *Mean and 95% CI*	Duration of MV(days) *Mean and 95% CI*
eICU	Day 1 (*1,549*)	***Berlin definition***	***Mild***	*14.27 [12.52, 16.01]*	*6.28 [6.06, 6.49]*
	Day 2 (*1,098*)			*16.21 [14.03, 18.39]*	*6.75 [6.47, 7.03]*
	Day 3 (*872*)			*16.39 [13.94, 18.86]*	*7.58 [7.25, 7.91]*
	Day 1 (*2,361*)		***Moderate***	*18.68 [17.11, 20.25]*	*6.98 [6.78, 7.19]*
	Day 2 (*1,454*)			*23.52 [21.33, 25.70]*	*7.94 [7.65, 8.24]*
	Day 3 (*1,025*)			*23.51 [20.91, 26.11]*	*8.75 [8.40, 9.10]*
	Day 1 (*1,243*)		***Severe***	*28.72 [26.20, 31.24]*	*8.30 [7.96, 8.65]*
	Day 2 (*429*)			*43.12 [38.42, 47.83]*	*9.09 [8.45, 9.72]*
	Day 3 (*429*)			*43.12 [38.42, 47.83]*	*9.09 [8.45, 9.72]*
	Day 1 (*1,309*)	***New severity criteria***	***Mild***	*14.36 [12.46, 16.27]*	*6.17 [5.94, 6.40]*
	Day 2 (*849*)			*15.31 [12.89, 17.74]*	*6.19 [5.89, 6.49]*
	Day 3 (*698*)			*14.61 [11.99, 17.24]*	*7.39 [7.04, 7.76]*
	Day 1 (*2,053*)		***Moderate***	*18.11 [16.45, 19.79]*	*6.72 [6.51, 6.94]*
	Day 2 (*1,182*)			*20.73 [18.41, 23.04]*	*7.64 [7.33, 7.95]*
	Day 3 (*822*)			*20.80 [18.02, 23.58]*	*8.18 [7.82, 8.53]*
	Day 1 (*1,791*)		***Severe***	*25.63 [23.60, 27.65]*	*8.18 [7.89, 8.47]*
	Day 2 (*950*)			*34.74 [31.70, 37.77]*	*9.01 [8.60, 9.43]*
	Day 3 (*806*)			*36.72 [33.39, 40.06]*	*9.42 [8.97, 9.87]*

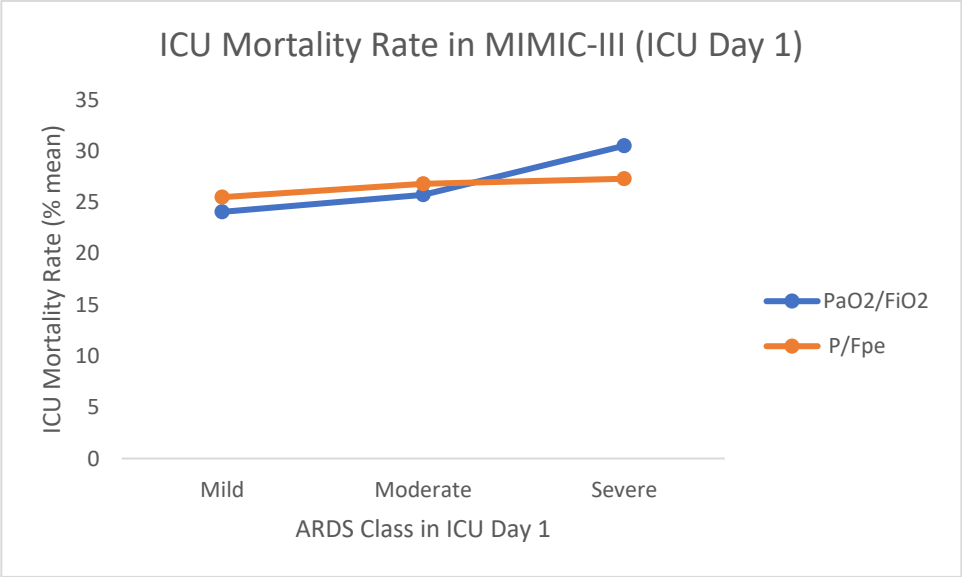


Figure A1. ICU mortality rate in MIMIC-III (ICU Day 1).

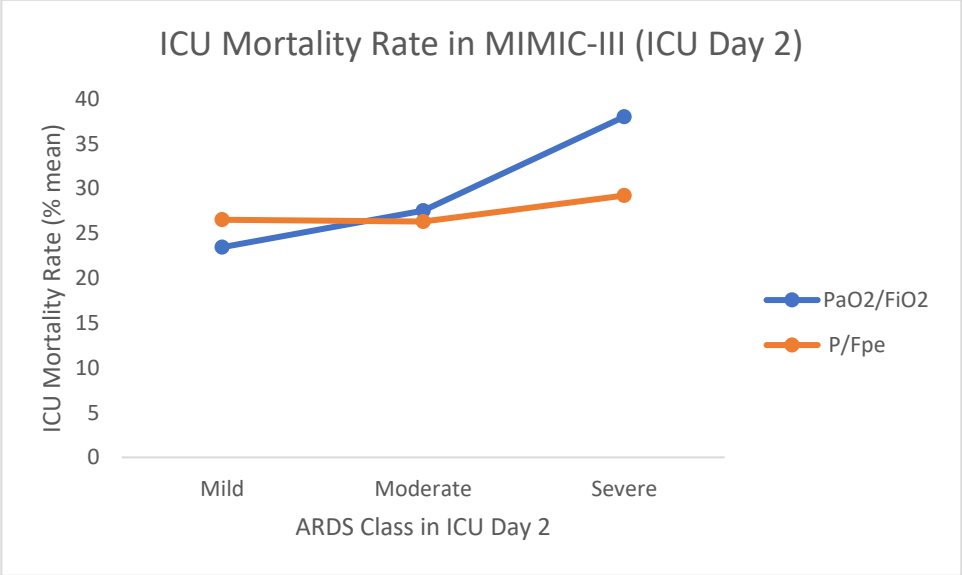


Figure A2. ICU mortality rate in MIMIC-III (ICU Day 2).

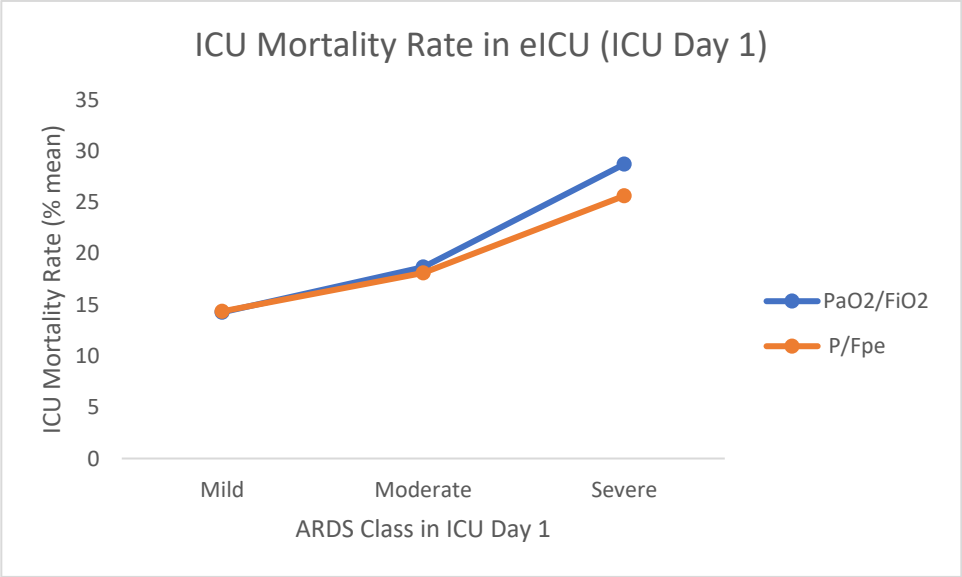


Figure A3. ICU mortality rate in eICU (ICU Day 1).

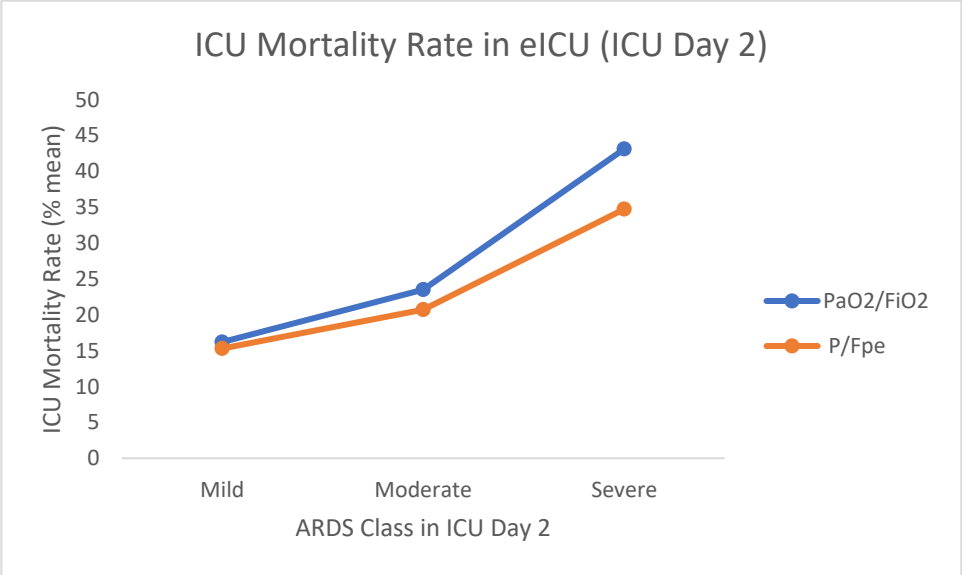


Figure A4. ICU mortality rate in eICU (ICU Day 2).

UNIVERSITAT
ROVIRA i VIRGILI