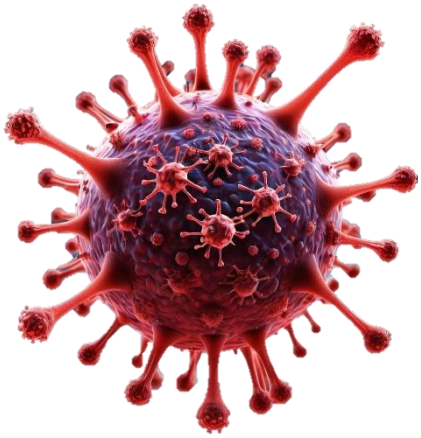

Traditional versus intentionally created severity scores for the prognosis of COVID-19 patients in Portugal

DANIELA ANDRADE MARQUES

(Licenciada em Tecnologia Biomédica)

Dissertação para obtenção do grau de Mestre em Engenharia Biomédica

Orientadores:

Doutora Iola Maria Silvério Pinto

Doutor Luís Bento

Júri:

Presidente:

Doutora Cecília Ribeiro da Cruz Calado

Vogais:

Doutora Alda Cristina Jesus Valentim Nunes de Carvalho

Doutora Iola Maria Silvério Pinto

Dezembro de 2024

Traditional versus intentionally created severity scores for the prognosis of COVID-19 patients in Portugal

DANIELA ANDRADE MARQUES

(Licenciada em Tecnologia Biomédica)

Dissertação para obtenção do grau de Mestre em Engenharia Biomédica

Orientadores:

Doutora Iola Maria Silvério Pinto (ISEL)

Doutor Luís Bento (CHULC)

Júri:

Presidente:

Doutora Cecília Ribeiro da Cruz Calado (ISEL)

Vogais:

Doutora Alda Cristina Jesus Valentim Nunes de Carvalho (UAb)

Doutora Iola Maria Silvério Pinto (ISEL)

Dezembro de 2024

This page was intentionally left blank

Acknowledgements

I would like to thank my supervisor Professor Doctor Iola Pinto for all the effort, dedication, motivation and demand to enrich my knowledge and skills in this whole project. I admire all the motivation that it was given to me to work better and to believe in myself. I am also grateful to Doctor Luís Bento for all the availability and teaching that were given, he has been always helpful throughout the journey. To Cristiana, for all the knowledge and for being always helpful.

To my friends and family for always supporting me unconditionally, always believing in my capabilities and for being proud of me.

Resumo

Contexto - A pandemia de COVID-19 motivou o desenvolvimento de ferramentas de prognóstico para pacientes em estado crítico. O cálculo e análise de *scores* de severidade tradicionais e *scores* criados especificamente para COVID-19 permite avaliar a capacidade discriminativa destes na tomada de decisão em contexto clínico e previsão da mortalidade de pacientes com COVID-19 alocados em unidades de cuidados intensivos (UCI).

Objetivo - Comparação do desempenho de *scores* de severidade tradicionais em UCI (APACHE II, SAPS II, SAPS 3, SOFA) com os *scores* desenvolvidos especificamente para COVID-19 (Shang-COVID, SEIMC, BURDEN e *Inflammation-based*), na previsão de quatro eventos relativos à mortalidade dos pacientes admitidos na UCI, durante a primeira e segunda vagas da pandemia em Portugal.

Métodos – Foram analisados dados de pacientes adultos com COVID-19 admitidos na UCI do *Centro Hospitalar Universitário Lisboa Central* em Lisboa, Portugal. Foram calculados oito *scores* de severidade para cada paciente e considerados quatro eventos: mortalidade hospitalar, mortalidade na UCI, mortalidade precoce na UCI (morte até 7 dias da admissão, inclusive) e mortalidade tardia na UCI (morte após 7 dias da admissão). A capacidade discriminativa dos *scores* foi avaliada através das estimativas das áreas abaixo da curva ROC, respetivas estimativas intervalares e valores-p.

Resultados – Para a mortalidade hospitalar o SEIMC teve o melhor desempenho, com AUCs de 0,810 (primeira vaga) e 0,723 (segunda vaga). O APACHE II e SAPS 3 também apresentaram bons valores de AUCs (>0,7), enquanto o BURDEN e INFLAMMATION-BASED apresentaram AUCs abaixo de 0,6. Para a mortalidade na UCI, o SEIMC destacou-se com AUC de 0,808 na primeira vaga. O SAPS 3 e APACHE II também apresentaram uma boa capacidade discriminativa (> 0,7). Na segunda vaga, o SEIMC apresentou uma AUC de 0,705. Para a morte precoce na UCI, o SAPS 3 foi o mais eficaz na segunda vaga (AUC de 0,828), seguido pelo SEIMC, enquanto o BURDEN e o *Inflammation-based* mostraram baixa capacidade discriminativa. Para a mortalidade tardia na UCI, o SEIMC e SAPS 3 destacaram-se na primeira vaga, tendo a capacidade discriminativa diminuído na segunda.

Conclusão - Embora os *scores* tradicionais sejam relevantes para o prognóstico dos pacientes com COVID-19, o novo *score* SEIMC mostrou uma capacidade discriminativa mais consistente para os quatro eventos. O bom desempenho dos *scores* tradicionais mostra que podem ser credíveis neste contexto, o que destaca a importância de validar continuamente estes *scores* para aprimorar o cuidado ao paciente e a alocação de recursos.

Palavras-chave - COVID-19 ● *Scores* ● Prognóstico ● UCI ● Curvas ROC

Abstract

Context - The COVID-19 pandemic motivated the development of prognostic tools for critically ill patients. The calculation and analysis of traditional severity scores and scores created specifically for COVID-19 allow evaluating their discriminatory capacity in decision-making in a clinical context and predicting mortality in patients with COVID-19 allocated to intensive care units (ICU).

Purpose - The goal was to compare the performance of traditional ICU severity scores (APACHE II, SAPS II, SAPS 3, SOFA) with scores developed specifically for COVID-19 (Shang-COVID, SEIMC, BURDEN, and inflammation-based), in predicting four mortality outcomes in ICU patients during the first and second waves of the pandemic in Portugal.

Methods - Data from adult patients with COVID-19 admitted to the ICU at the Centro Hospitalar Universitário Lisboa Central in Lisbon, Portugal, were analysed. Eight severity scores were calculated for each patient, and four outcomes were considered: hospital mortality, ICU mortality, early ICU mortality (death within 7 days of admission), and late ICU mortality (death after 7 days of admission). The discriminative ability of the scores was evaluated through ROC curve area estimates, their respective confidence intervals, and p-values.

Results - For hospital mortality, SEIMC performed best with AUCs of 0.810 (first wave) and 0.723 (second wave). APACHE II and SAPS 3 also showed good AUCs (>0.7), while BURDEN and Inflammation-base had AUCs below 0.6. For ICU mortality, SEIMC stood out with an AUC of 0.808 in the first wave. SAPS 3 and APACHE II also demonstrated good discriminative ability (>0.7). In the second wave, SEIMC had an AUC of 0.705. For early ICU death, SAPS 3 was the most effective in the second wave (AUC of 0.828), followed by SEIMC, while BURDEN and INFLAMMATION-BASED showed low discriminative ability. For late ICU mortality, SEIMC and SAPS 3 stood out in the first wave, with discriminative ability decreasing in the second wave.

Conclusion - While traditional scores are relevant for predicting outcomes in COVID-19 patients, the newly developed SEIMC score demonstrated more consistent discriminative ability across all four outcomes. The good performance of traditional scores indicates that traditional scores can still be reliable in this context, emphasizing the importance of continuously validating severity scores to improve patient care and resource allocation.

Keywords - COVID-19 ● Scores ● Prognosis ● ICU ● ROC Curves

This page was intentionally left blank

Table of Contents

List of figures

List of tables

List of abbreviations

APACHE II	Acute Physiology and Chronic Health Evaluation II
AUC	Area Under the Curve
COVID-19	Coronavirus Disease 2019
CRP	C-Reactive Protein
ECMO	Extracorporeal Membrane Oxygenation
ECMO	Extracorporeal Membrane Oxygenation
eGFR	Estimated Glomerular Filtration Rate
ICU	Intensive Care Unit
IMV	Invasive Mechanical Ventilation
IQR	Interquartile Range
LDH	Lactate Dehydrogenase
O2Hb	Oxyhemoglobin
PO2	Partial Pressure of Oxygen
ROC	Receiver Operating Characteristic
SAPS 3	Simplified Acute Physiology Score 3
SAPS II	Simplified Acute Physiology Score II
SARS-CoV	Severe Acute Respiratory Syndrome Coronavirus
SOFA	Sequential Organ Failure Assessment

This page was intentionally left blank

1. Objectives and work structure

The COVID-19 pandemic has had a profound effect on healthcare systems globally, with cases ranging from asymptomatic to severe, including pneumonia, acute respiratory failure, and acute respiratory distress syndrome (ARDS). Critically ill patients often require intensive care, with treatment options spanning oxygen therapy, non-invasive ventilation, and invasive mechanical ventilation, (Chung et al., 2023).

Severity scoring systems are vital tools in healthcare, providing numerical measures to classify disease progression and predict outcomes, such as mortality or intensive care unit (ICU) admission. These scores allow clinicians to prioritize resources and optimize patient flow, helping manage the surge in cases during crises like COVID-19. To develop these systems, statistical methods such as logistic regression analysis and principal components analysis are employed, selecting and weighting clinical and laboratory variables to produce an overall score that indicates disease severity, (Rapsang & Shyam, 2014).

The COVID-19 pandemic, triggered by the SARS-CoV-2 virus, has underscored the importance of early identification of patients at high risk for clinical deterioration, essential for managing healthcare resources and delivering timely care (Siddiqui et al., 2022). Traditional scoring systems have been adapted and new models developed to meet the specific challenges of COVID-19. Various studies aim to assess the predictive accuracy of these scoring systems in forecasting mortality and ICU admission rates among hospitalized COVID-19 patients, (Martin et al., 2022).

In this study, (Martin et al., 2022), scores such as the Sequential Organ Failure Assessment (SOFA), the Simplified Acute Physiology Score II (SAPS II), the Simplified Acute Physiology Score 3 (SAPS 3), the Acute Physiology and Chronic Health Evaluation II (APACHE II), the SEIMC score, the Shang, Burden and Inflammation-Based scores are included in the evaluation of COVID-19 severity.

Severity scores have proven effective in gauging disease progression. For instance, age and comorbidity-inclusive scores demonstrate enhanced accuracy in predicting mortality among COVID-19 patients, while hypoxemia-based scores, such as SOFA, better predict ICU admissions. This demonstrates the need for tailored scoring strategies to refine risk stratification and optimize resource use, (Martin et al., 2022).

To explore whether the performance of these scoring systems changed as the pandemic progressed, it was selected two waves of COVID-19 cases for comparison. The scores were calculated from scratch using clinical data from CHULC and the formulas were the ones presented in the literature. All patients were different from first wave to second wave with no duplicate cases between these waves. This approach allows us to assess whether the

discriminative capacity of these scores has evolved over time, providing a benchmark to determine consistency and reliability in different phases of the pandemic.

To accomplish the primary goal outlined above, the study also aims to:

- Assess the predictive accuracy of both traditional (e.g., APACHE II, SAPS II, SAPS 3) and COVID-19-specific scores (SEIMC, Shang, BURDEN and Inflammation-based) in determining patient outcomes, such as mortality and ICU admission;
- Identify which scores demonstrate the highest reliability for COVID-19 cases and evaluate their effectiveness;
- Investigate if the discriminative power of various severity scores shifted between the first two waves of COVID-19, reflecting possible changes in disease management, patient characteristics, or healthcare response;
- Provide a comparative benchmark to assess whether the performance of these scores improved or remained consistent over time;
- Determine which scoring systems offer the best guidance for patient stratification, helping to direct ICU admissions and optimize resource allocation;
- Support clinical decision-making by identifying scores that accurately reflect patient risk profiles, allowing for more effective patient flow management;
- Provide recommendations for the development of modified or new scoring systems specifically geared toward pandemic-related healthcare challenges;
- Compare mortality outcomes (hospital mortality, ICU mortality, early and late ICU mortality) to determine which scores aligns best with the observed outcomes, thereby validating the best tools for ICU decision-making;
- Contribute to critical care improvement by identifying scoring tools that offer high discriminative power and accuracy across various patient outcomes, especially for high-risk patients.

Comparative studies, analysing tools like SAPS II, SAPS 3, and APACHE II, reveal moderate accuracy in predicting ICU mortality, though traditional models may underestimate mortality in COVID-19 cases, (Martin et al., 2022). Scoring systems that include inflammation-based and thrombotic considerations, like the SEIMC and Shang Burden scores, provide additional insights into disease severity, accommodating COVID-19's unique inflammatory response.

This study will examine the association of various severity scoring systems with four outcomes: hospital mortality, ICU mortality, early ICU mortality (within 7 days), and late ICU mortality (after 7 days). Utilizing raw data from the Intensive Care Department at Central Lisbon University Hospital Center (CHULC), the goal is to calculate and determine which scores best

predict mortality outcomes. By analyzing both traditional and COVID-19-specific scoring systems across two pandemic waves, this research will enable a comparative assessment to reveal any shifts in predictive accuracy. This approach aims to enhance risk stratification and guide clinical decision-making, ultimately improving patient care and outcomes during the COVID-19 pandemic and in similar future crises.

2. Literature review

2.1. Brief introduction of SARS-Cov-2 pandemic

In mid-December 2019, a novel coronavirus was identified in patients experiencing respiratory illnesses in Wuhan, China. Due to its symptom similarity to SARS and genomic resemblance, it was designated SARS-CoV-2. By March 16, 2020, the virus had spread to 114 countries, infecting 167,515 people and causing 6,606 deaths, prompting WHO to declare the first coronavirus-induced pandemic on March 11, 2020. As the outbreak expanded, numerous viral strains were isolated and sequenced, with over 500 complete or near-complete genomes available by March 18, 2020. SARS-CoV-2 is the seventh known human coronavirus, alongside SARS-CoV, MERS-CoV, and four milder strains (229E, HKU1, NL63, OC43). Evidence suggests all human coronaviruses originated from animals, primarily bats and rodents. While SARS-CoV-2 shares structural similarities with other coronaviruses, its sequence differs significantly from human-infecting betacoronaviruses like SARS-CoV, MERS-CoV, and HKU-1. However, it shows 96% similarity to a bat coronavirus from Yunnan, China, indicating a likely bat origin, (Chaw et al., 2020).

Throughout the global SARS-CoV-2 pandemic, five key mutated strains have emerged: Alpha, Beta, Gamma, Delta, and Omicron. The first significant mutation, D614G in the spike protein (S protein), spread globally in July 2020. Two months later, the Alpha variant was identified in the United Kingdom (Davies et al., 2021; Hart et al., 2022). In December 2020, the Beta variant was found in South Africa, followed by the Gamma variant in Brazil in January 2021, the Delta variant in the UK in March 2021, and the Omicron variant in Botswana in November 2021. As of July 10, 2022, more than 551 million cases had been confirmed across over 194 countries, with over 6 million deaths, and daily infections continued to increase at a significant pace, as shown in Figure 2.1.1, (Hao et al., 2022).

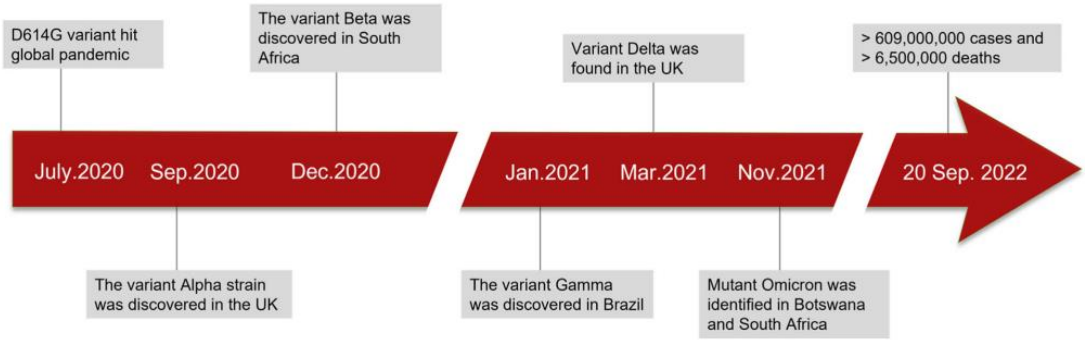

Figure 2.1.1 - Timeline of the development of SARS-CoV-2 variants, (Hao et al., 2022).

SARS-CoV-2 is classified as a Sarbecovirus within the Betacoronavirus genus of the coronavirus family. Its single-stranded RNA genome encodes 4 structural proteins (spike, membrane, envelope, and nucleocapsid), 16 non-structural proteins, and 9 accessory factors (Figure 2.1.2). The 5' end contains two open reading frames (ORF1a and ORF1b) that produce polyproteins pp1a and pp1ab, which are cleaved into 16 non-structural proteins by viral proteases nsp3 (PLpro) and nsp5 (3CLpro/Mpro). These non-structural proteins are crucial for viral replication, transcription, translation, and genome modification. Key conserved proteins like nsp3, nsp5, nsp12 (RdRp), and the nucleocapsid protein are potential antiviral drug targets. The RdRp gene, essential for RNA virus replication, serves as a stable genetic marker for evolutionary analysis due to its high conservation across RNA viruses, (Hao et al., 2022).

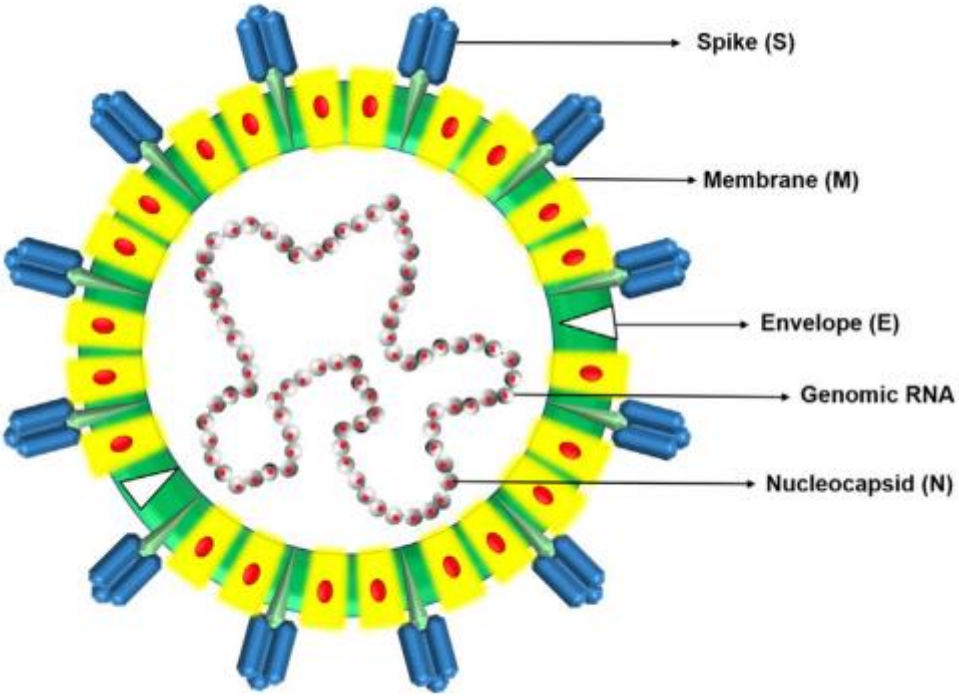

Figure 2.1.2 - Illustration of a typical coronavirus structure (ranging from 80 to 120 nanometers in diameter), depicting its key structural components: the spike (S), membrane (M), and envelope (E) proteins on the surface, with the nucleocapsid (N) protein encasing, (Hao et al., 2022).

The first confirmed cases of COVID-19 in Portugal were recorded in March 2020. Table 2.1.1 shows the number of confirmed cases up to March 3, 2020, in Portugal compared to the rest of the world. There were also 101 notifications of suspected cases in Portugal up to that date, with suspected cases beginning to emerge in February. This information highlights the early stages of the COVID-19 pandemic in Portugal. The country's first confirmed cases appeared in early March, while suspected cases had been reported since February. This

timeline aligns with the global spread of the virus, as many European countries began seeing their first cases around this time. The comparison between Portugal and the rest of the world in Table 1 likely illustrates the relatively low number of confirmed cases in Portugal at that point compared to the global total. This reflects the initial phase of the pandemic in the country, before widespread community transmission had taken hold, ("Direção Geral de Saúde", 2020).

Table 2.1.1 - Comparison of COVID-19 cases: World and Portugal, ("Direção Geral de Saúde", 2020).

Metric	World	Portugal
Confirmed cases	90 663	4
Deceased	3 043	0
Suspected cases notified	Not provided	101 (since January 2020)
Active community transmission	China (Mainland and Hong Kong), Iran, Italy (Emilia-Romagna, Lombardy, Piedmont, Veneto), Japan, Singapore, South Korea	Not mentioned

Figure 2.1.3 shows the number of COVID-19 cases in Portugal from March 2020 to September 2024, ("NÚMERO DE NOVOS CASOS E ÓBITOS POR DIA - Covid 19", 2024). A peak in the number of cases is observed between December 2021 and March 2022. The state of alert in Portugal ended on September 30, 2022, with the reduction in the number of cases. This description provides an overview of Portugal's COVID-19 situation over a four-year period. The end of the state of alert in September 2022 indicates that the situation had improved significantly by that time ("Fim Do Estado de Alerta - XXIII Governo - República Portuguesa", 2022).

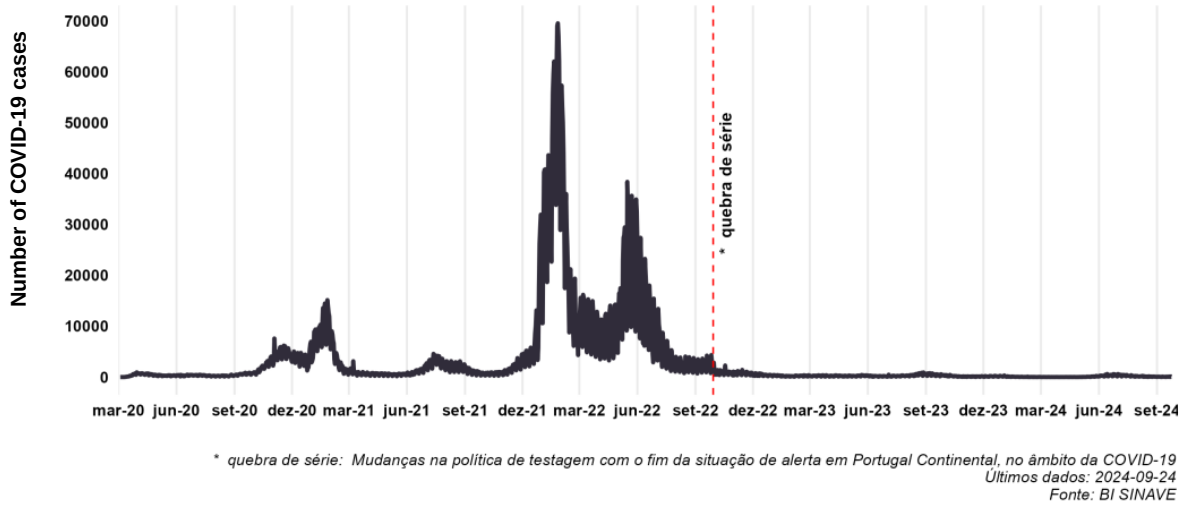

Figure 2.1.3 - COVID-19 cases in Portugal from March 2020 to September 2024, ("NÚMERO DE NOVOS CASOS E ÓBITOS POR DIA - Covid 19", 2024).

The first deaths associated with COVID-19 in Portugal occurred in March 2020, with a peak in death cases between December 2020 and March 2021, as shown in figure 2.1.4, ("NÚMERO DE NOVOS CASOS E ÓBITOS POR DIA - Covid 19", 2024).

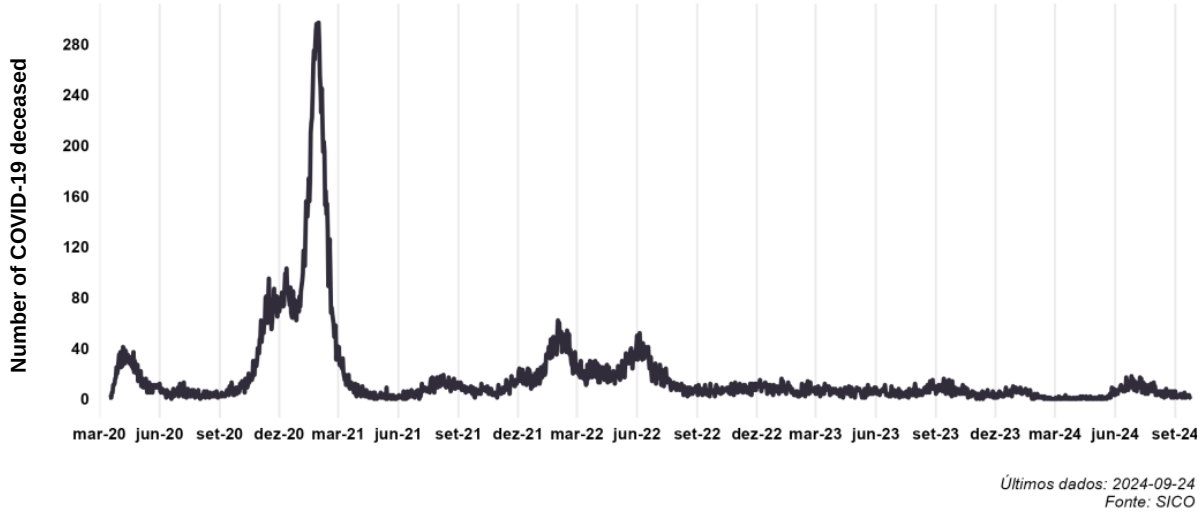

Figure 2.1.4 - COVID-19 deceased in Portugal from March 2020 to September 2024, ("NÚMERO DE NOVOS CASOS E ÓBITOS POR DIA - Covid 19", 2024).

Determining how SARS-CoV-2 first emerged is not only a matter of historical interest but is crucial for the development of future global public health strategies. By understanding the virus's origin, the scientific community and policymakers can better prepare for and mitigate the risk of future pandemics, potentially preventing similar outbreaks from taking such a devastating toll on humanity.

2.2. Age and gender

In the context of COVID-19, susceptibility to infection and symptom presentation varies significantly across age groups. Studies indicate that individuals under 20 years of age are approximately half as susceptible to COVID-19 when compared to those aged 20 and older. This age-related difference in susceptibility has implications for both transmission dynamics and public health strategies. Additionally, early symptoms of COVID-19 appear to differ between younger and older populations. While younger adults (ages 16-59) tend to present with more typical symptoms such as fever, cough, and loss of smell or taste, older adults (60+ years) often experience atypical or fewer symptoms, which can complicate timely diagnosis. This atypical presentation in older adults, combined with potential delays in seeking medical care, poses challenges in early detection and treatment, emphasizing the need for heightened awareness and tailored healthcare approaches for this vulnerable population. Understanding these variations is critical for improving diagnostic protocols and optimizing care for different age groups, (Unim et al., 2021), ("Early COVID-19 Symptoms Differ among Age Groups, Research Finds | King’s College London", 2021).

Figure 2.2.1 represents the cases, total hospitalizations, ICU admissions, and deceased in Portugal from the first to the seventh of November 2022, can be observed, which corresponds to the report available on the DGS website, ("DGS", 2022),

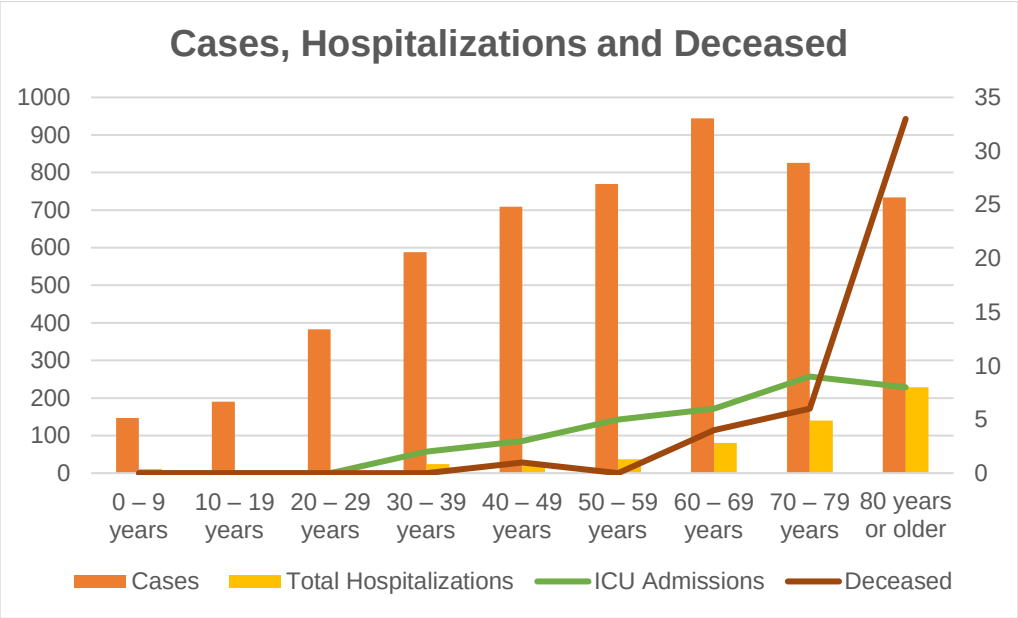

Figure 2.2.1 - COVID-19 cases, total hospitalizations, ICU admissions and deceased in Portugal from the first to the seventh of November 2022, ("DGS", 2022).

2.3. Main comorbidities

A research analysed death certificates in Portugal where COVID-19 was listed as the primary cause of death, using data from the country's National e-Death Certificates Information System (SICO platform), ("SICO – SPMS", 2013). The study period spanned from March 16 to December 31, 2020. To assess comorbidities, the researchers employed the Charlson and Elixhauser indexes, which are standardized methods for categorizing coexisting medical conditions based on ICD-10 diagnostic codes. For population-based calculations, the study utilized official estimates of Portugal's resident population for the year 2020, (Nogueira et al., 2022).

Comorbidities present (identified according to the Elixhauser comorbidity index) total, by sex can be observed in Figure 2.3.1. Hypertension (uncomplicated) emerged as the most prevalent condition, affecting 30% of the total population. This was followed by uncomplicated diabetes (16.83%) and congestive heart failure (14.48%). The least prevalent condition was blood loss anemia, affecting only 0.03% of the population. Congestive heart failure demonstrated the most pronounced gender disparity, with a prevalence of 16.85% in males compared to 12.31% in females. Hypertension also showed a slight male predominance (31.01% vs. 29.07%), (Nogueira et al., 2022).

Diabetes exhibited a marginally higher prevalence in males (17.35%) compared to females (16.36%). Similarly, obesity was more prevalent in males (5.85%) than females (4.68%). Solid tumors without metastasis showed a higher prevalence in females (8.05%)

compared to males (5.97%). Chronic pulmonary disease was slightly more prevalent in females (7.48%) than males (6.75%). Neurological disorders were more common in females (5.28%) compared to males (4.41%), (Nogueira et al., 2022).

Certain conditions demonstrated minimal gender differences: cardiac arrhythmias (10.57% in males, 10.68% in females), renal failure (12.25% in males, 13.36% in females), fluid and electrolyte disorders (4.00% in males, 3.68% in females), blood loss anemia (0.03% in both genders), (Nogueira et al., 2022).

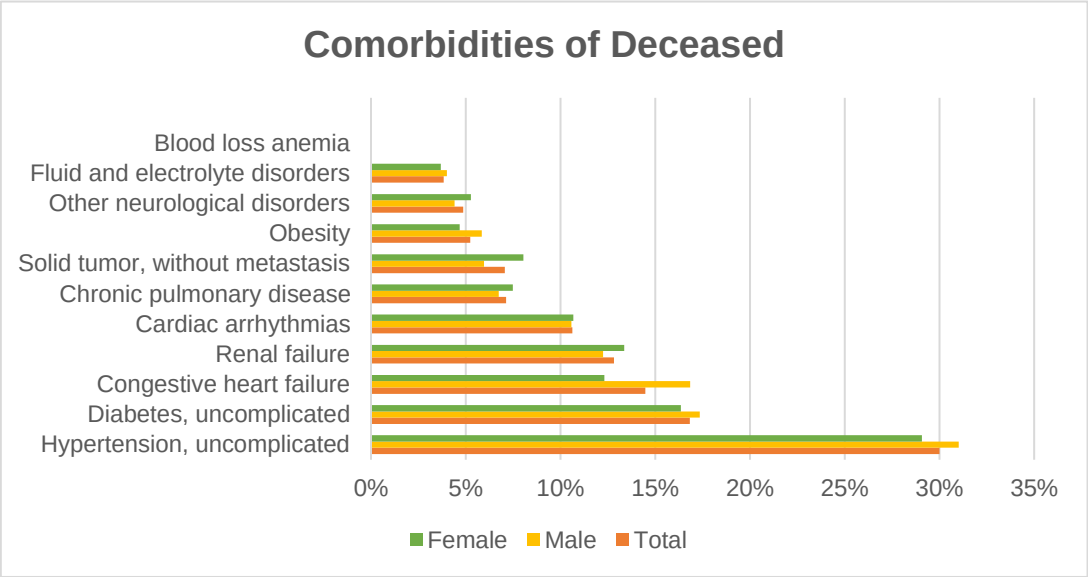

Figure 2.3.1 - Comorbidities present (identified according to the Elixhauser comorbidity index), total, by sex, (Nogueira et al., 2022).

2.4. Most common symptoms

COVID-19, caused by the SARS-CoV-2 virus, presents with a broad range of symptoms that vary significantly in severity across individuals. While many infected individuals display characteristic signs of respiratory illness, others may experience gastrointestinal or neurological symptoms or may remain entirely asymptomatic. This variation in symptomatology, coupled with the potential for severe outcomes, underscores the necessity of early detection and medical intervention, particularly among high-risk populations.

COVID-19, caused by the SARS-CoV-2 virus, presents a wide range of symptoms that can vary in severity. This overview aims to provide a detailed examination of the signs and symptoms associated with COVID-19 infection, as well as its incubation period, infectious period, complications, and high-risk groups. The most common symptoms of COVID-19 include fever (temperature ≥ 38.0°C or 100.4°F), persistent cough, fatigue, shortness of breath

or difficulty breathing, loss or alteration of smell (anosmia or hyposmia), and loss or disturbance of taste (ageusia or dysgeusia). Additional symptoms that may occur include muscle aches and pains, sore throat, nasal congestion or runny nose, headache, and gastrointestinal issues such as nausea, vomiting, and diarrhea, ("COVID-19", 2023).

The incubation period for COVID-19 typically ranges from 2 to 14 days, with an average of 4 to 5 days. Individuals may be infectious approximately 48 hours before the onset of symptoms. The duration of infectiousness can vary depending on viral variants, disease severity, and individual immunity, but generally extends up to 10 days after symptom onset, or up to 20 days in cases of chronic illness or severe immunosuppression. In severe cases, COVID-19 can lead to serious complications, including post-COVID-19 condition (long COVID), acute respiratory failure with severe pneumonia, cardiac failure and multi-organ dysfunction, and death. Certain populations are at higher risk of developing severe complications from COVID-19, including individuals aged 60 years and older, pregnant women, people with severe immunosuppression, and those with chronic medical conditions, ("COVID-19", 2023).

Table 2.4.1 delineates the prevalence of various symptoms across three common respiratory illnesses: COVID-19, Influenza (Flu), and the Common Cold. This comparative analysis serves to elucidate the symptomatic similarities and differences among these conditions, which is crucial for both clinical differentiation and public health awareness. The data reveals a substantial overlap in symptomatology between COVID-19 and Influenza, with both conditions sharing several common symptoms. This similarity poses a significant challenge in clinical differentiation based solely on symptomatic presentation, ("Constipação, Gripe e COVID-19 | Hospital Da Luz", 2020).

COVID-19 demonstrates a unique symptom profile with the common occurrence of ageusia and anosmia (loss of taste or smell), which is rare in both Influenza and the Common Cold. This distinctive feature may serve as a potential diagnostic indicator for COVID-19, although it is not universally present in all cases. Influenza is characterized by a higher prevalence of chills compared to COVID-19 and the Common Cold, potentially offering a subtle differentiating factor. The Common Cold, in contrast, is distinguished by the common occurrence of sneezing, a symptom that is rare in both COVID-19 and Influenza, ("Constipação, Gripe e COVID-19 | Hospital Da Luz", 2020).

All three conditions present with respiratory symptoms, albeit with varying frequencies. Cough is uniformly common across all three illnesses. However, dyspnea (shortness of breath) is more frequently associated with COVID-19 compared to Influenza and the Common Cold. Rhinorrhea (runny nose) and pharyngitis (sore throat) show higher prevalence in COVID-19 and the Common Cold relative to Influenza, ("Constipação, Gripe e COVID-19 | Hospital Da Luz", 2020).

Systemic manifestations such as fever, fatigue, cephalalgia (headache), and myalgia (muscle aches) are predominantly associated with COVID-19 and Influenza. Their occurrence in the Common Cold is comparatively less frequent, providing a potential basis for preliminary differentiation. Gastrointestinal symptoms, including nausea, vomiting, and diarrhea, are occasionally observed in both COVID-19 and Influenza but are rarely associated with the Common Cold. This pattern suggests a broader systemic involvement in COVID-19 and Influenza compared to the primarily upper respiratory tract-focused Common Cold, ("Constipação, Gripe e COVID-19 | Hospital Da Luz", 2020).

The substantial symptom overlap, particularly between COVID-19 and Influenza, underscores the critical importance of laboratory diagnostics in achieving accurate differentiation. The reliance on clinical presentation alone for definitive diagnosis is fraught with potential for misclassification, highlighting the necessity for confirmatory testing, especially in distinguishing COVID-19 from other respiratory illnesses.

Table 2.4.1 - Prevalence of various symptoms across three common respiratory illnesses: COVID-19, Influenza (Flu), and the Common Cold, ("Constipação, Gripe e COVID-19 | Hospital Da Luz", 2020).

Symptom	COVID-19	Influenza (Flu)	Common Cold
Fever	Common	Common	Rare
Cough	Common	Common	Common
Fatigue	Common	Common	Sometimes
Shortness of breath	Common	Sometimes	Rare
Loss of taste or smell	Common	Rare	Rare
Runny or stuffy nose	Common	Sometimes	Common
Sore throat	Common	Sometimes	Common
Headache	Common	Common	Rare
Muscle or body aches	Common	Common	Sometimes
Nausea or vomiting	Sometimes	Sometimes	Rare
Diarrhea	Sometimes	Sometimes	Rare
Sneezing	Rare	Rare	Common
Chills	Sometimes	Common	Rare

2.5. Severity scores

When describing severity scores such as APACHE II, SAPS II, and SAPS 3, it is essential to emphasize that these scores were originally validated for hospital mortality. These scores were developed to assess the severity of illness in hospitalized patients and estimate the risk of death during hospitalization. Their applicability in clinical practice has been rigorously tested in various populations, and the validity of these scores has been widely confirmed in critical patients. This means that risk calculations based on these scores are highly accurate in predicting in-hospital mortality. Higher score values indicate greater severity of the patients' health status. However, when used outside of this context (such as predicting long-term mortality or in non-hospital settings), their precision may be different, which should be considered when applying them to broader studies.

Acute physiology and chronic health evaluation (APACHE) is the main scoring system used in intensive care units. The first version evaluated the severity of disease with 34 physiological parameters, and it was presented by Knaus in 1981. The next version, APACHE II, calculates the risk of hospital death and it was released in 1985, (Rapsang & Shyam, 2014).

APACHE II score is composed by three parts: a) Twelve acute physiological parameters, also known as the acute physiology score (APS); b) the patient's age, and c) chronic diseases and prior surgical procedures (Rapsang & Shyam, 2014). The APACHE II score, the primary diagnostic category used for ICU admission, and whether emergency surgery was necessary for the patient are used to predict hospital mortality. With the use of the logistic regression equation and specially designed beta coefficients, the predicted probability of hospital death is determined, (Rapsang & Shyam, 2014).

The final APACHE II score is derived from the summation of scores assigned to physiological parameter like consciousness level, body temperature, mean arterial pressure, heart rate, respiratory rate, alveolar-capillary gradient, pH level, serum concentrations of HCO3-, Na+, K+, and creatinine, as well as leukocyte and haematocrit counts. These measurements are collected within the initial 24 hours of the patient's admission to the intensive care unit (ICU), with the values furthest from the baseline (normal) being selected for the final calculations, (Rapsang & Shyam, 2014).

The APACHE II scoring system incorporates several pre-existing health conditions, including liver cirrhosis, advanced heart failure (NYHA class IV), chronic obstructive pulmonary disease (COPD), chronic renal failure necessitating dialysis, and immune deficiency. These conditions are typically diagnosed and recorded prior to the patient's hospital admission. Furthermore, an additional 2 points are assigned to patients with any of these chronic diseases, those who are immunocompromised, or those who underwent elective surgeries before ICU

admission. Non-surgical patients or those admitted for emergency postoperative care receive 5 points, (Rapsang & Shyam, 2014).

Initially, it was decided that the APACHE II score would not be utilized for patients under the age of 16, those admitted to the ICU for less than 8 hours, individuals with extensive burns, or those who underwent coronary artery bypass grafting. Additionally, patients who were readmitted to the ICU during the same hospital stay were excluded from the database cohorts. A study involving 5815 ICU patients demonstrated a notable correlation between the APACHE II score and mortality rates, (Rapsang & Shyam, 2014).

The APACHE II score is described in Table 2.5.1, with the variables outlined and their corresponding scores. After assigning the score for each variable, the scores are summed to determine the score value, (Rapsang & Shyam, 2014).

Furthermore, time bias exists because the physiological variables are all dynamic and subject to a variety of influences, including ongoing resuscitation and treatment. This is a crucial factor to consider when treating patients in the intensive care unit, particularly considering the recent emphasis on the significance of early goal-directed therapies. Each of these factors carries the potential to cause an overestimation of anticipated mortality, (Rapsang & Shyam, 2014).

Table 2.5.1 - APACHE II score. APACHE II is equal to the sum of A + B + C sections, (Rapsang & Shyam, 2014).

Component	Variable	High Abnormal Range				Normal Range	Low Abnormal Range			
		4	3	2	1	0	1	2	3	4
A: Acute Physiological Score	Temperature rectal (°C)	≥41	39-40.9	-	38.5-38.9	36-38.4	34-35.9	32-33.9	30-31.9	≤29.0
	Mean arterial pressure (mm Hg)	≥160	130-159	110-129	-	70-109	-	50-69	-	≤49
	Heart rate-ventricular response	≥180	140-179	110-139	-	70-109	-	55-69	40-54	≤39
	Respiratory rate per minute (non-ventilated or ventilated)	≥50	35-49	-	25-34	12-24	10-11	6-9	-	≤5
	Oxygenation: A-aDO2 or PaO2 (Torr)	≥500	350-499	200-349	-	<200	-	-	-	-
	FiO2 ≥0.5 record A-aDO2	-	-	-	-	PO2>70	PO2 61-70	-	PO2 55-60	PO2<55
	FiO2 <0.5 record only PaO2	-	-	-	-	-	-	-	-	-
	Arterial pH	≥7.7	7.6-7.69	-	7.5-7.59	7.33-7.49	-	7.25-7.32	7.15-7.24	<7.15
	Serum HCO3 (mmol/L) (only if no ABGs)	≥52	41-51.9	-	32-40.9	23-31.9	-	18-21.9	15-17.9	<15
	Serum sodium (mmol/L)	≥180	160-179	155-159	150-154	130-149	-	120-129	111-119	≤110
	Serum potassium (mmol/L)	≥7	6-6.9	-	5.5-5.9	3.5-5.4	3-3.4	2.5-2.9	-	≤2.5
	Serum creatinine (µmol/L)	≥350	200-340	150-190	-	60-140	-	<60	-	-
	Hematocrit (%)	≥60	-	50-59.9	46-49.9	30-45.9	-	20-29.9	-	<20
	White blood cell count (×1,000/mm³)	≥40	-	20-39.9	15-19.9	3-14.9	-	1-2.9	-	<1
	Glasgow Coma Score (GCS)	15 minus actual GCS								

B: Age Points	**Age (years)**		**Points**
	≤44		0
	45-54		2
	55-64		3
	65-74		5
	≥75		6
C: Chronic Health Points	**History**	**Points for elective surgery**	**Points for emergency surgery**
	Liver: Biopsy-proven cirrhosis and documented portal hypertension or prior episodes of hepatic failure	2	5
	Cardiovascular: NYHA Class IV	2	5
	Respiratory: e.g., severe COPD, hypercapnia, home O2, pulmonary hypertension	2	5
	Immunocompromised	2	5
	Renal: Chronic dialysis	2	5

Most of the variables used in APACHE II are also utilized in APACHE III, however the method for gathering neurological data has changed—the GCS is no longer employed. It adds two crucial elements in particular: the lead-time bias and the patient's origin. One diagnosis must be chosen, and the acute diagnosis is considered. The APACHE III scores range from 0 to 299 points, with 252 points allocated to the 18 physiological variables, 24 points for age, and 23 points for the chronic health status. These variables are selected to enhance the model's explanatory capacity. The scores are calculated as the most abnormal values from the first 24 hours spent in the intensive care unit, (Rapsang & Shyam, 2014).

Initially introduced in 1993 by Le Gall et al., SAPS II serve as a tool for assessing the severity of patients in the intensive care unit (ICU). This model incorporates a total of 17 variables, comprising 12 physiological parameters, age, type of admission, and three disease-related factors. Like other scoring systems, SAPS II records the most severe values of the selected variables within the first 24 hours following admission. The SAPS II score ranges from 0 to 163 points, with allocations of 0-116 points for physiological variables, 0-17 points for age, and 0-30 points for pre-existing diagnoses. To estimate the probability of death, logistic regression analysis was employed (Sakr et al., 2008). However, it's worth noting that the SAPS II model's discrimination and calibration may not align perfectly when applied to a new patient population, (Sakr et al., 2008).

This score is detailed in Table 2.5.2, with the variables and corresponding scores according to the pattern in which the values of each patient variable are found, whether they present any chronic disease, and the type of admission to the ICU.

At the end of the Table 2.5.2., the mortality risk percentage is presented according to the SAPS II score value. For example, a score of 29 corresponds to a 10% mortality risk, while a score of 77 corresponds to a 90% mortality risk. Thus, it is considered that the higher the score value, the higher the percentage of mortality risk.

Table 2.5.2 - SAPS II score, (Sakr et al., 2008).

Variable	Range	Points
Age	<40	0
	40-59	7
	60-69	12
	70-74	15
	75-79	16
	≥80	18
Heart rate	<40	11
	40-69	2
	70-119	0
	120-159	4
	≥160	7
Systolic BP	<70	13
	70-99	5
	100-199	0
	≥200	2
Body temperature	<39°C	0
	≥39°C	3
PaO2/FiO2 ratio (if ventilated or CPAP)	<100	11
	100-199	9
	≥200	6
Urinary output	<0.5 L/day	11
	0.5-0.999 L/day	4
	≥1 L/day	0
Serum urea	<28 mg/dL	0
	28-83 mg/dL	6
	≥84 mg/dL	10
WBC count	<1,000/mm³	12
	1,000-19,000/mm³	0
	≥20,000/mm³	3
Serum potassium	<3 mEq/L	3
	3-4.9 mEq/L	0
	≥5 mEq/L	3
Serum sodium	<125 mEq/L	5
	125-144 mEq/L	0
	≥145 mEq/L	1

Serum bicarbonate	<15 mEq/L	6
	15-19 mEq/L	3
	≥20 mEq/L	0
Bilirubin	<4 mg/dL	0
	4-5.9 mg/dL	4
	≥6 mg/dL	9
Glasgow Coma Scale	<6	26
	6-8	13
	9-10	7
	11-13	5
	14-15	0
Chronic diseases	Metastatic cancer	9
	Hematologic malignancy	10
	AIDS	17
Type of admission	Scheduled surgical	0
	Medical	6
	Unscheduled surgical	8

Twenty distinct factors that are simple to measure at patient admission (within the first hour) form the basis of SAPS 3. It enables early risk assessment, separating patient status from the standard of treatment in the intensive care unit. Following a thorough application of cross-validation methods, the SAPS III score demonstrated extremely strong internal validity. However, prospective validation in distinct demographics and with better defined ICU patients enhances the model's generalizability, or suitability for diverse contexts, (Rapsang & Shyam, 2014).

The SAPS 3 prognostic model, designed to predict hospital mortality for patients admitted to intensive care units (ICUs), employs a comprehensive scoring system based on 20 variables. These variables, collected within one hour before or after ICU admission, encompass patient characteristics, circumstances of admission, and physiological parameters. The scoring table delineates the specific criteria and point allocations for each variable, providing insight into the model's structure and the relative importance assigned to different factors in predicting mortality risk, (Rapsang & Shyam, 2014)

The SAPS 3 score (Table 2.5.3) is derived from three main categories of variables. First, Patient Characteristics and Comorbidities includes age (0-18 points), comorbidities (0-8 points), and the use of major therapeutic options before ICU admission (0-3 points). The age variable demonstrates a non-linear progression, with points increasing more rapidly for older age groups, reflecting the exponential rise in mortality risk with advancing age, (Moreno et al., 2005).

Second, the Circumstances of ICU Admission encompasses factors like admission type (0-8 points), ICU admission source (0-7 points), length of stay before ICU admission (0-7 points), and infection at ICU admission (0-5 points). The distinction between scheduled

surgical, unscheduled surgical, and medical admissions highlights how different admission circumstances affect patient outcomes, (Moreno et al., 2005).

Lastly, Physiological Parameters includes a broad array of variables such as the Glasgow Coma Scale (0-15 points), total bilirubin (0-5 points), body temperature (0-7 points), creatinine (0-8 points), heart rate (0-7 points), leukocytes (0-2 points), pH (0-3 points), platelets (0-13 points), systolic blood pressure (0-11 points), and oxygenation (0-11 points). These variables allow for a thorough assessment of multi-organ dysfunction, (Moreno et al., 2005).

The SAPS 3 model assigns points to each variable based on predefined criteria, with the total score being the sum of these points. The non-linear point allocation observed in certain variables, such as age and creatinine, reflects the intricate relationships between these factors and mortality risk. The oxygenation variable is particularly noteworthy, as it differentiates between ventilated and non-ventilated patients, recognizing the substantial impact of mechanical ventilation on outcomes and the need for distinct assessment criteria based on ventilation status, (Moreno et al., 2005).

Moreover, the model's inclusion of pre-ICU factors, such as comorbidities and the length of hospital stay before ICU admission, is a key feature. This approach acknowledges that a patient's condition upon ICU admission is significantly shaped by their pre-existing health status and the events leading up to their ICU care, (Moreno et al., 2005).

Table 2.5.3 - SAPS 3 score, (Moreno et al., 2005).

Variable	Criteria	Points
Age	<40	0
	40-59	5
	60-69	9
	70-74	13
	75-79	15
	≥80	18
Comorbidities	None	0
	Cancer therapy	3
	Hematologic cancer	6
	AIDS	8
Use of major therapeutic options before ICU admission	No	0
	Yes	3
Admission	Scheduled surgical	0

	Unscheduled surgical	6
	Medical	8
ICU admission source	Operating room	0
	Emergency room	5
	Other	7
Length of stay before ICU admission	<14 days	0
	14-27 days	6
	≥28 days	7
Infection at ICU admission	No	0
	Nosocomial	4
	Respiratory	5
Glasgow Coma Scale	3-4	15
	5	10
	6	7
	7-12	2
	>=13	0
Total bilirubin (mg/dL)	<2	0
	2-5.9	4
	≥6	5
Body temperature (°C)	<35	7
	≥35	0
Creatinine (mg/dL)	<1.2	0
	1.2-1.9	2
	2-3.4	7
	≥3.5	8
Heart rate (bpm)	<120	0
	120-159	5
	≥160	7
Leukocytes (x10³/mm³)	<15	0
	≥15	2
pH	≥7.25	0
	<7.25	3
Platelets (x10³/mm³)	<20	13
	20-49	8
	50-99	5
	≥100	0
Systolic blood pressure (mmHg)	<40	11
	40-69	8
	70-119	3
	≥120	0
Oxygenation	If ventilated or CPAP:	
	PaO2/FiO2 <100	11
	PaO2/FiO2 100-199	9
	PaO2/FiO2 ≥200	6

If not ventilated:	
PaO2 <60	5
PaO2 ≥60	0

The SOFA system, established through a consensus meeting of the European Society of Intensive Care Medicine in 1994 and subsequently refined in 1996, serves as a crucial tool for assessing the severity of illness in critically ill patients. Vincent et al. conducted a comprehensive evaluation of the SOFA subjective score in 1998, analyzing data from 1449 patients. This scoring system quantifies the degree of organ dysfunction across six organ systems, assigning scores ranging from 0 to 4. Notably, the presence of one organ failure, combined with respiratory failure, indicates the lowest mortality risk, while other combinations are associated with mortality rates between 65% and 74%, (Rapsang & Shyam, 2014).

Further analyses have explored the prognostic value of maximal SOFA scores and their changes over time. It has been observed that while the maximal score provides significant prognostic information, changes in the SOFA score over time exhibit a lesser predictive value. Additionally, the temporal evolution of the patient's condition throughout their ICU stay is considered in prognostic assessments, (Rapsang & Shyam, 2014).

While direct conversion of SOFA scores to mortality rates is not feasible, rough estimates of mortality risk can be inferred from prospective studies. Sequential evaluations of organ dysfunction during the initial days of ICU admission serve as reliable indicators of prognosis. Studies by Bale et al. and Ferreira et al. demonstrated the predictive utility of mean and maximal SOFA scores, with high scores at 48 hours predicting increased mortality rates, (Rapsang & Shyam, 2014).

SOFA score is presented in Table 2.5.4. Despite the fact that the SOFA score cannot be directly converted to mortality, two published prospective studies can be used to approximate the risk of death, (Rapsang & Shyam, 2014).

Table 2.5.4 - SOFA score, (Rapsang & Shyam, 2014).

Sequential organ failure assessment score						
Organ System	**Score**					
	Variable	**0**	**1**	**2**	**3**	**4**
Pulmonary	Lowest PaO2 (Torr)/FiO2 (%)	>400	≤400	≤300	≤200+respiratory support	≤100+respiratory support
Coagulation	Lowest platelet (103/mm3)	>150	≤150	≤100	≤50	≤20
Hepatic	Highest bilirubin (µmol/L)	<20	20-32	33-101	102-204	>204
Circulatory	Blood pressure status	Mean arterial pressure (mmHg) >70	Mean arterial pressure (mmHg) <70	Dopamine* dose≤5 or dobutamine any dose	Dopamine dose>5 or epinephrine≤0.1 or norepinephrine≤0.1	Dopamine dose>15 or epinephrine>0.1 or norepinephrine>0.1
Neurologic	GCS	15	13-14	10-12	6-9	<6
Renal	Highest creatinine level (µmol/L)	<110	110-170	171-299	300-440	>440
	Total urine output (mL/24 h)				<500	<200
Score	0-6	7-9	10-12	13-14	15	15-24
Score %	<10	15-20	15-20	50-60	>80	>90

In order to enable doctors to appropriately evaluate the medical resources of patients with varying risks, the goal of this study was to identify risk indicators for death, in severe ill patients and develop a risk model to predict mortality, as shown in Table 2.5.5, (Shang et al., 2020).

Three institutions were involved in this retrospective cohort study: Leishenshan Hospital (Wuhan, China), No. 7 Hospital of Wuhan, and Zhon-gnan Hospital of Wuhan University. The government-assigned COVID-19 treatments are handled by Zhongnan Hospital. Since January 2020, Zhongnan Hospital of Wuhan University has entrusted No. 7 Hospital of Wuhan, one of the designated institutions for the hospitalization of patients with COVID-19, (Shang et al., 2020).

The definition of fever was considered at 37.3 °C or more at the axilla. The Third International Consensus Definition from 2016 was used to describe septic shock. The definition of acute kidney injury (AKI) was established by kidney disease: Improving Global Outcomes. If bilirubin or liver enzymes were more than twice the upper limit of normal, it was considered acute liver injury (ALI). The diagnosis of acute myocardial injury (AMI) was made if new abnormalities were seen in electrocardiography and echocardiography, or if the serum levels of cardiac biomarkers were higher than the upper 99th percentile limit, (Shang et al., 2020).

The Chinese version 7.0 of the COVID-19 management guidelines was used to define the sickness severity. The COVID-19 patient population is divided into four categories: mild, which refers to mild clinical symptoms without an imaging feature of pneumonia; ordinary; this category includes clinical symptoms like fever and cough with an imaging feature of pneumonia; severe; this category includes dyspnoea; respiratory frequency <30/min; blood oxygen saturation <93%; partial pressure of arterial oxygen to fraction of inspired oxygen ratio < 300; and/or lung infiltrates > 50% within 24 to 48 hours; and critical ill cases. The 7th Edition's categorization of severe and critical illness was used to the examination of severe cases. From the moment of sickness onset to death, survival time was determined, (Shang et al., 2020).

Table 2.5.5 describes Shang scoring system. Low risk scoring is considered equal or less than 2 points and high-risk scoring is considered bigger than 2.

Table 2.5.5 - Shang COVID score, (Shang et al., 2020).

	Criteria	**Score**
Age	<60 years	0
	60-75 years	1
	>75 years	2

Coronary heart disease		1
Lymphocytes	<8%	1
Procalcitonin	>0.15 ng/mL	2
D-dimer	>500 ng/mL	1

The goal of the SEIMC score study was to create and verify a mortality prediction model for COVID-19 patients visiting emergency rooms in hospitals. The databases of two sizable retrospective cohorts of COVID-19 hospitalized patients in Spain in 2020 served as the data source. The derivation cohort, funded by the Spanish Society of Infectious Diseases and Clinical Microbiology (SEIMC), was a multicenter cohort of patients hospitalized from 2 February to 17 March. The follow-up censoring date was set for April 17, (Berenguer et al., 2021).

The SEIMC score for predicting 30-day mortality in patients with COVID-19 is defined using a formula that combines various predictors such as age, SaO2, Neutrophil-to-Lymphocyte Ratio, eGFR, Dyspnoea, and Sex, with regression coefficients obtained from the predictive model. This score was developed by dividing each coefficient by the coefficient with the lowest value and rounding to an integer, creating risk groups based on the 30-day probability of death according to the simplified score, (Berenguer et al., 2021).

Table 2.5.6 describes the SEIMC score with the several variables defined. This score can vary in a range from 0 to 30 points. It is considered a low age-adjusted saturation of oxygen on room air, defined as ≤90% for patients aged >50 years and ≤93% for patients aged ≤50 years. If the saturation of oxygen is low, considering these defined values, it is assigned 2 points, but if the saturation of oxygen is not considered to be low it is assigned 0 points, (Berenguer et al., 2021).

The outcome was 30-day all-cause mortality, measured from the day of the hospital admission. Patients that were discharged alive before 30 days after admission were assumed to have survived for at least 30 days, (Berenguer et al., 2021).

Table 2.5.6 - SEIMC score, (Berenguer et al., 2021). Lower risk: 0-2 points; moderate risk: 3-5 points; high risk: 6-8 points and very high risk: 9-30 points.

COVID-19 SEIMC Score

Risk factor	Addition to risk score	Risk score
Age (years)		
<40	0	
40-54	1	

55-64	3	
65-74	5	
75-79	9	
80-84	14	
85-89	15	
≥90	21	
Low age-adjusted SaO2*		
No	0	
Yes	2	
Neutrophil-to-lymphocyte ratio		
<3.22	0	
3.22-6.33	1	
>6.33	2	
eGFR mL/min/1.73 m2 (CKD-EPI)		
≥60	0	
30-59	2	
<30	3	
Dyspnea		
No	0	
Yes	1	
Sex		
Female	0	
Male	1	
	Total risk score	**0 to 30**

*≤90% for patients aged >50 years and ≤93% for patients aged ≤50 years

The final model of COVID-19 BURDEN score was approximate the values of the coefficients for each category by creating an integer number using the coefficients produced by the logistic regression equation to develop a numerical scoring model. The O2 Sat scores of more than 90%, 84–90%, and less than 84% were determined using the regression analyses' coefficients, and they were assigned 0, 1, and 2, respectively. For additional variables, such as CRP, PT, DBP, BUN, and LDH levels, scores of 0 or 1 were assigned (Table 2.5.7), (Imanieh et al., 2023).

The burden score is calculated by assigning points to six clinical variables (Table 2.5.7), each of which reflects different aspects of a patient's condition. First, C-reactive protein (CRP) levels, a marker of inflammation, are evaluated. If CRP is ≤73.1 mg/L, 0 points are assigned, indicating low inflammation. If CRP is >73.1 mg/L, 1 point is assigned, reflecting higher inflammation. Next, oxygen saturation variation (O2 saturation variation) is considered, where

a variation greater than 90% results in 0 points, suggesting normal respiratory function. A variation between 84-90% is awarded 1 point, indicating moderate concern, while a variation less than 84% receives 2 points, indicating significant respiratory impairment, (Imanieh et al., 2023).

Prothrombin time (PT), a measure of blood clotting, is also assessed. A PT of ≤16.2 seconds scores 0 points, indicating normal clotting function, while a PT >16.2 seconds scores 1 point, suggesting impaired coagulation. Diastolic blood pressure (DBP) is then evaluated, where a DBP greater than 75 mmHg results in 0 points, indicating normal cardiovascular function, and a DBP ≤75 mmHg scores 1 point, reflecting possible hypotension, (Imanieh et al., 2023).

Blood urea nitrogen (BUN), which assesses kidney function, is another variable. If BUN is ≤23 mg/dL, 0 points are awarded, indicating normal kidney function, whereas BUN >23 mg/dL results in 1 point, indicating potential kidney impairment. Finally, lactate dehydrogenase (LDH) levels, which can indicate tissue damage, are measured. If LDH is ≤731 U/L, 0 points are assigned, indicating normal levels, while LDH >731 U/L scores 1 point, suggesting tissue damage or stress, (Imanieh et al., 2023).

The total burden score is the sum of the points from these six variables, with higher scores indicating a greater overall burden of illness. This scoring system helps clinicians evaluate the severity of a patient's condition and guide treatment decisions accordingly, (Imanieh et al., 2023).

Table 2.5.7 - COVID-19 BURDEN score, (Imanieh et al., 2023).

Variables	Points awarded
C-reactive protein (CRP) mg/L	
≤73.1 mg/L	0
>73.1 mg/L	1
O2 saturation variation (O2 sat Variation)	
Greater than 90%	0
84-90%	1
Less than 84%	2
Increased prothrombin time (increased PT)	
≤16.2 s	0
>16.2 s	1
Diastolic blood pressure (DBP) mmHG	

>75 mmHg	0
≤75 mmHg	1
Blood urea nitrogen (BUN) mg/dL	
≤23 mg/dL	0
>23 mg/dL	1
Raised lactate dehydrogenase (Raised LDH) (U/L)	
≤731 Units/lit (U/L)	0
>731 Units/lit (U/L)	1

The inflammation-based risk score (Table 2.5.8) is based on three inflammatory biomarkers: C-reactive protein (CRP), leukocytes count, and serum albumin. Each biomarker was assigned weighted points based on its association with in-hospital mortality: 1 point for WBC count ≥9.3 × 10^3 cells/μL, 2 points for CRP level ≥13.0 mg/L, and 3 points for serum albumin level ≤3.6 g/dL. The total score categorizes systemic inflammation into four levels: no signs (0 points), mild (1-2 points), moderate (3-4 points), and severe (5-6 points). This approach allowed for a standardized assessment of inflammation severity in COVID-19 patients based on established biomarkers and cutoff values, (Amezcua-Guerra et al., 2021).

Table 2.5.8 - Inflammation-based score points and inflammation levels, (Amezcua-Guerra et al., 2021).

Parameters		**Points**
Leukocytes	≥9.3 × 10³/μL	1
CRP	≥13.0 mg/L	2
Albumin	≤3.6 g/dL	3
Total Score	**Inflammation Level**	
0	No systemic inflammation	
1-2	Mild inflammation	
3-4	Moderate inflammation	
5-6	Severe inflammation	

2.6. Comparison of COVID-19 scores

Table 2.6.1 compares four severity scores (SHANG, SEIMC, BURDEN, and Inflammation-based) based on their variables and calculation methods, each focusing on different aspects of patient assessment.

The SHANG score uses a broad range of variables, including age, coronary heart disease, lymphocytes, procalcitonin, D-dimer, C-reactive protein (CRP), oxygen saturation variation, albumin, prothrombin time, diastolic blood pressure, neutrophil-to-lymphocyte ratio, eGFR, dyspnea, and sex. These variables are scored based on specific thresholds, with the total score ranging from 0 to 30. This comprehensive approach considers both clinical and biochemical factors, allowing for a nuanced prediction of severity, especially in COVID-19 patients.

The SEIMC score, on the other hand, focuses on fewer parameters, including CRP, procalcitonin, D-dimer, lymphocytes, age, sex, blood urea nitrogen, albumin, and neutrophil-to-lymphocyte ratio. Each of these variables is awarded points based on predefined thresholds. While the SEIMC score is simpler and more focused, it remains comprehensive in assessing the severity of COVID-19 by emphasizing inflammatory markers and organ function.

The BURDEN score is like the SHANG score but differs in the weight assigned to each variable. It includes age, coronary heart disease, lymphocytes, procalcitonin, D-dimer, CRP, oxygen saturation variation, prothrombin time, diastolic blood pressure, neutrophil-to-lymphocyte ratio, eGFR, dyspnea, and sex. The distinct scoring system and variable weights allow for a balanced assessment of severity, but the BURDEN score may have different predictive power based on the importance placed on each parameter.

The INFLAMMATION-BASED score focuses primarily on markers of inflammation, such as CRP and lactate dehydrogenase, along with clinical factors like blood urea nitrogen, prothrombin time, oxygen saturation variation, age, and sex. This score is simpler compared to the others, with a particular emphasis on the inflammatory response, which is crucial in assessing severe COVID-19 cases.

In summary, these severity scores differ in the range and focus of the variables they incorporate. SHANG and BURDEN offer a more extensive evaluation of clinical and biochemical parameters, while SEIMC and INFLAMMATION-BASED focus more specifically on inflammation and key clinical signs. The choice of score depends on the need for a comprehensive assessment or a more focused evaluation of specific risk factors.

Table 2.6.1 - Comparison of SEIMC, SHANG, BURDEN and Inflammation-based scores.

Score	Variables	Calculation
SHANG	Age, Coronary heart disease, Lymphocytes, Procalcitonin, D-dimer, C-reactive protein, Oxygen saturation variation, Albumin, Prothrombin time, Diastolic blood pressure, Neutrophil-to-lymphocyte ratio, eGFR, Dyspnea, Sex	Scored based on age and various clinical parameters, with total score ranging from 0 to 30.
SEIMC	C-reactive protein, Procalcitonin, D-dimer, Lymphocytes, Age, Sex, Blood urea nitrogen, Albumin, Neutrophil-to-lymphocyte ratio	Each variable is awarded points based on thresholds. Total score is calculated from these points.
BURDEN	Age, Coronary heart disease, Lymphocytes, Procalcitonin, D-dimer, C-reactive protein, Oxygen saturation variation, Prothrombin time, Diastolic blood pressure, Neutrophil-to-lymphocyte ratio, eGFR, Dyspnea, Sex	Like SHANG, but with different weight for each parameter and specific scoring system.
INFLAMMATION-BASED	C-reactive protein, Lactate dehydrogenase, Blood urea nitrogen, Prothrombin time, Diastolic blood pressure, Oxygen saturation variation, Age, Sex	Variables related to inflammation (CRP, LDH) and clinical factors (Oxygen saturation, Prothrombin time) determine total score.

3. Methodologies

3.1. Study population and data assembly

This study focuses on COVID-19 patients admitted to the Intensive Care Unit (ICU) at São José Hospital, part of *Centro Hospitalar Lisboa Central* (CHULC) from March 10 to August 22, 2020 (first wave) and from August 23 to December 19, 2020 (second wave), considering both the first and second waves in accordance with the hospital's guidelines. The research is a component of the "Predictive Models of COVID-19 Outcomes for Higher Risk Patients Towards Precision Medicine" (PREMO) project, which received ethical approval from the institution's ethics board and complies with all relevant legal and ethical guidelines.

The reason for studying the two waves of COVID-19 goes beyond simply monitoring the progression of a viral disease. Over time, the virus changed, both in terms of its transmissibility and virulence (with the emergence of new variants). Additionally, therapeutic approaches were progressively adjusted, moving from experimental treatments to more effective and evidence-based strategies. Furthermore, many COVID-specific scores were initially validated with populations from the first wave, making it important to reassess them in subsequent waves to see if they are still applicable or if the changes in population characteristics and treatment strategies have affected their effectiveness. Therefore, studying the different waves helps to better understand the changes in patient profiles, treatments, and the impact of viral variants, ultimately aiding in the adjustment and improvement of severity scores to reflect the current reality. All scores were calculated from scratch using clinical data from CHULC and the formulas used were the ones presented in literature.

A database was created using Microsoft Excel to compile demographic information, details on symptom onset, and the dates of hospital and ICU admissions and discharges. This data was collected from the hospital’s electronic medical record system. In conducting this master's research, the study population was restricted to adult participants aged 18 and above. Furthermore, only individuals with complete and adequate data pertaining to the key variables under investigation were included in the final analysis. This selection criteria ensured that the dataset was both relevant to the adult population and sufficiently comprehensive to support robust findings.

Table 3.1.1. presents the comorbidities that were collected from the data.

Table 3.1.1 – List of comorbidities.

Category	Comorbidities
Cardiovascular Diseases	Arterial Hypertension
	Ischemic heart disease
	Congestive heart failure
	Pulmonary Hypertension
	Cardiac Dysrhythmia
Metabolic Disorders	Diabetes
	Dyslipidemia
	Hyperuricemia
	Obesity
	Hypothyroidism
Respiratory Diseases	Chronic respiratory disease
Renal Diseases	Chronic kidney disease
Hematologic and Oncologic Conditions	Solid Cancer
	Hematologic cancer
	Amyloidosis
Neurological Disorders	Multiple Sclerosis
	Parkinsons
	Epilepsy
	Schizophrenia
Infectious and Autoimmune Diseases	AIDS
	Autoimmune disease

Other Conditions	HBP
	Depression
	Chronic liver disease
	Stroke
	History of organ transplant

3.2. Clinical data

Table 3.2.1 presents various clinical parameters alongside their respective values of interest, specifying whether the minimum, maximum, or both values are relevant for evaluation.

Oxyhemoglobin (O2Hb), arterial oxygen saturation (So2), base excess/deficit (BE), bicarbonate (HCO3-), partial pressure of oxygen (Po2), lymphocytes (count and percentage), and erythrocytes all require monitoring of their minimum values. This indicates that low levels of these parameters could be of clinical concern. On the other hand, parameters such as platelets (expressed as count or percentage) also require monitoring of their minimum values, while lactate dehydrogenase (LDH) and alanine transaminase (ALT) have their maximum values of interest, suggesting that elevated levels could signal pathology, (Von Rekowski, 2022).

Similarly, albumin is monitored for its minimum value, whereas low-density lipoprotein (LDL) cholesterol, high-density lipoprotein (HDL) cholesterol, and C-reactive protein (CRP) are evaluated based on their maximum values, reflecting that higher values could indicate increased risk or inflammatory states. Additionally, blood urea, creatine kinase (CK), creatinine, D-dimers, and procalcitonin are also measured for their maximum values, where elevated levels may indicate potential issues such as kidney dysfunction, muscle damage, or infection/inflammation, (Von Rekowski, 2022).

Lastly, neutrophils (both count and percentage) require monitoring of both their maximum and minimum values, emphasizing that both high and low levels could be clinically significant, depending on the patient's condition, (Von Rekowski, 2022).

Table 3.2.1 - List of variables used (adapted from (Von Rekowski, 2022)).

Parameter	Value of interest	Parameter	Value of interest
Oxyhemoglobin (O2Hb)	Minimum	Platelets (x 10^9/L) / (%)	Minimum
Arterial oxygen saturation (So2)	Minimum	LDH (U/L)	Maximum
Base excess/deficit (BE (ecf))	Minimum	Albumin (g/L)	Minimum
Bicarbonate ($HCO3^-$(act))	Minimum	ALT (U/L)	Maximum
Partial pressure of oxygen (Po2)	Minimum	LDL cholesterol (mg/dL)	Maximum
Lymphocytes (x 10^9/L) / (%)	Minimum	HDL cholesterol (mg/dL)	Maximum
Blood Urea (mg/dL)	Maximum	C-Reactive Protein (CRP) (mg/L)	Maximum
Creatine kinase (CK) (U/L)	Maximum	Neutrophiles (x 10^9/L) / (%)	Maximum and minimum
Creatinine (mg/dL)	Maximum	D-Dimers (µg/L)	Maximum
Erythrocytes	Minimum	Procalcitonin (ng/mL)	Maximum

3.3 Statistical Analysis

Categorical variables were represented by their absolute frequencies (percentages) and continuous variables as mean (standard deviation) or median (25th-75th percentile), when these were presented with an asymmetric distribution and deviations from normality. To assess the normality of the continuous variable's, Kolmogorov-Smirnov and Shapiro-Wilk tests were used, as appropriate. Additionally, given the objectives and according to clinical criteria, some variables were categorized, and others were recoded. To access the discriminative power for the severity scores, concerning death outcomes, ROC curve was obtained, and AUC estimate is reported with its 95% confidence interval.

Descriptive and inferential statistics were obtained by SPSS software (version 26.0). A level of significance of 0.05 was considered.

The effectiveness of a binary diagnostic classification system is assessed analytically using a graph known as the ROC curve. One of the well-defined dichotomous categories, such as the existence or absence of a disease, needs to be applied to the diagnostic test results. It is necessary to establish a reference value, or cut-off value, for diagnosis because a lot of test findings are displayed as continuous or ordinal variables. Thus, the cut-off value can be used to detect whether a disease is present, (Nahm, 2022).

The ROC curve seeks to determine the ideal cut-off value with the optimum diagnostic performance as well as to categorize a patient's illness condition as either positive or negative

based on test results. The ROC curve can also be used to compare the results of two or more tests and assess a test's overall diagnostic performance, (Nahm, 2022).

The two main categories of ROC curve types are parametric and nonparametric (also known as empirical). Table 3.3.1 lists the benefits and drawbacks of these two approaches. The binary method is another name for the parametric technique. The parametric ROC curve assumes the shape of a smooth curve by increasing the sample size and linking an infinite number of points. When two independent groups with differing means and standard deviations either follow a normal distribution or satisfy the normality assumption via algebraic conversion or square root transformation, this method uses a maximum likelihood estimation to estimate the curve, (Nahm, 2022).

Table 3.3.1 - Nonparametric and parametric ROC curve, (Nahm, 2022).

	Nonparametric ROC curve	**Parametric ROC curve**
Pros	No need for assumptions about the distribution of data. Provides unbiased estimates of sensitivity and specificity. The plot passes through all points. Uses all data. Computation is simple.	Shows a smooth curve. Compares plots at any sensitivity and specificity value.
Cons	Has a jagged ir staircase appearance. Compares plots only at observed values of sensitivity or specificity.	Actual data are discarded. Curve does not necessarily go through actual points. ROC curves and AUC are possibly biased. Computation is complex.

AUC is a commonly used metric to assess how accurate diagnostic tests are. The test's accuracy increases with the ROC curve's proximity to the upper left corner of the graph, where the sensitivity and false positive rate are both equal to one (specificity = 1). AUC = 1.0 is hence the perfect ROC curve, (Nahm, 2022).

Nonetheless, a graph is created on the 45° diagonal (y = x) of the ROC curve (AUC = 0.5) when the coordinates of the x-axis (1 – specificity) and the y-axis correspond to 1: 1 (i.e., true positive rate = false positive rate). In this case, using a coin toss or other incidental approach to determine the presence or absence of disease makes no sense as a diagnostic tool. Thus, the AUC needs to be more than 0.5 for any diagnostic method to have relevance, and it should generally be more than 0.8 to be deemed appropriate. Furthermore, the ROC curve with the

greatest AUC is thought to have a better diagnostic performance when comparing the results of two or more diagnostic tests, (Nahm, 2022).

The test performed uses the null hypothesis (H_0) that there is no difference in the AUCs (Area Under the Curve) between each pair of scores. Therefore, when the p-value is less than 0.05, we reject the null hypothesis (H_0) and conclude that there is a statistically significant difference between the respective AUCs (Equation 3.3.1). ΔAUC denotes the difference between the two AUC values. AUC_1 is the AUC of the first score. AUC_2 is the AUC of the second score.

$$\Delta AUC = AUC_1 - AUC_2 \tag{3.3.1}$$

For independent samples, the Z-Test by Delong et al. (1988) was considered and for paired samples the Z-Test by Zhou et al. (2022) was considered, ("NCSS, LLC.", 2021). These tests are for AUCs comparison.

With the aim of comparing the discriminatory capacity of each score in relation to the two waves of COVID-19, a test was carried out to compare AUCs in independent samples, (Nahm, 2022).

3.4. Sensitivity, specificity, false positive and false negative

Understanding the concept of sensitivity and specificity, which are used to assess a diagnostic test's performance, is a prerequisite to comprehending the ROC curve. Sensitivity is the ability of a test to correctly identify patients with a disease and specificity is the ability of a test to correctly identify people without the disease. According to Table 3.4.1, TP are the true positives, where the test is positive and the disease is confirmed, FN are false negatives, where the test is negative but the disease is present, FP is false positives and TN is true negatives. A test with a sensitivity and specificity of 1.0 would be excellent, but it is rare condition in medical practice, (Nahm, 2022).

Table 3.4.1 – Contingency table, (Nahm, 2022).

		Predicted condition	
		Test (+)	**Test (-)**
True condition	Disease (+)	TP	FN
	Disease (-)	FP	TN

4. Results and Discussion

4.1. Results for First Wave Data

4.1.1. Clinical and demographic characteristics

Table 4.1.1.1 presents a comprehensive overview of patient characteristics, outcomes, and severity scores for a cohort of 125 COVID-19 patients admitted to the intensive care unit (ICU), during the first COVID-19 wave. The median age of the patients is 67 (P_{25} = 53.5; P_{75}= 76) years. The gender distribution shows a significant male predominance, with 99 (79.2%) of the patients being male and 26 (20.8%) females.

A vast majority of the patients, 114 (91.2%) had at least one comorbidity, while only 11 (8.8%) had no reported comorbidities. Regarding respiratory support, 100 (80%) of the patients required invasive mechanical ventilation (IMV), and 10 (8%) needed extracorporeal membrane oxygenation (ECMO), indicating the severity of their condition.

The outcomes data reveal that 27 (21.6%) patients died in the ICU, with 14 (11.2%) classified as early deaths and 13 (10.4%) as late deaths. The total hospital mortality count was 35 (28%). The median length of stay in the ICU was 9 days (P_{25} = 4.50; P_{75} = 15.0). The median APACHE II score was 15 (P_{25} = 11.0; P_{75}= 20.0) while the median SAPS II and SAPS 3 scores were 50 (P_{25} = 41.0; P_{75}= 57.0) and 59 (P_{25} = 50.0; P_{75}= 66.0) respectively. The median SOFA score was 7 (P_{25} = 4.0; P_{75}= 9.0). The Shang-COVID score had a median of 4 (P_{25} = 3.0; P_{75}= 5.0), and the SEIMC score's median was 9 (P_{25} = 6.50; P_{75}= 13.0). The BURDEN score, which had 3 missing values (2.4%), showed a median of 3 (P_{25} = 3.0; P_{75}= 4.0). The Inflammation-based score had a higher number of missing values, 30 (24%) and a median of 3 (P_{25} = 2.0; P_{75}= 3.0).

Table 4.1.1.1 - Clinical characteristics and demographics of first COVID-19 wave.

Variables	Missing count (%)	All Patients (N=125)
Age, years		67.0 (53.50 - 76.00)
Gender		
Male		99 (79.2)
Female		26 (20.8)
Presence of Comorbidities		
Yes		114 (91.2)
No		11 (8.8)
Respiratory support		
IMV		100 (80.0)
ECMO		10 (8.0)
Outcomes		
Total deceased at ICU		27 (21.6)

Early deceased at ICU		14 (11.2)
Late deceased at ICU		13 (10.4)
Total deceased at hospital		35 (28.0)
Days in ICU		9.0 (4.50-15.0)
Severity scores		
APACHE II		15.0 (11.0 - 20.0)
SAPS II		50.0 (41.0 - 57.0)
SAPS 3		59.0 (50.0 - 66.0)
SOFA		7.0 (4.0 - 9.0)
Shang-COVID		4.0 (3.0 - 5.0)
SEIMC		9.0 (6.50 - 13.0)
BURDEN	3 (2.4)	3.0 (3.0 - 4.0)
Inflammation-based	30 (24.0)	3.0 (2.0 - 3.0)

4.1.2. Statistical description for the severity scores for first wave

Table 4.1.2.1 summarizes various severity scores for a cohort of 125 patients in the context of COVID-19, reveals important insights into the severity of illness in this cohort. The APACHE II score, ranging from 4 to 38, shows a median of 15.0 (P_{25} = 11.0; P_{75}= 20.0), indicating moderate severity on average. The SAPS II and SAPS 3 scores demonstrate higher severity levels, with SAPS II ranging from 18 to 105 (median 50.0, (P_{25} = 41.0; P_{75}= 57.0) and SAPS 3 from 35 to 100 (median 59.0, (P_{25} = 50.0; P_{75}= 66.0)). The SOFA score, assessing organ dysfunction, ranges from 1 to 17 with a median of 7.0 (P_{25} = 4.0; P_{75}= 9.0), suggesting moderate organ dysfunction across the cohort. The COVID-specific Shang COVID score ranges from 1 to 7, with a median of 4.0 (P_{25} = 3.0; P_{75}= 5.0), indicating moderate severity specific to COVID-19. The SEIMC score, ranging from 2 to 23 with a median of 9.0 (P_{25} = 6.50; P_{75}= 13.0), suggests moderate to high severity. The BURDEN score, ranging from 1 to 7 with a median of 3.0 (P_{25} = 3.0; P_{75}= 4.0), and the Inflammation-based score, ranging from 1 to 3 with a median of 3.0 (P_{25} = 2.0; P_{75}= 3.0), both indicate moderate to high levels of severity and inflammation respectively. Notably, all scores show a p-value <0.05 for the normality test, indicating that none of them follow a normal distribution. This suggests that non-parametric statistical methods should be used when analysing these scores, and that the median and interquartile range may be more representative of the central tendency and spread than the mean and standard deviation. The non-normal distribution of these scores also reflects the heterogeneous nature of disease severity in this patient population, highlighting the complexity of managing critically ill COVID-19 patients.

Table 4.1.2.1 – Statistical description for severity scores.

Severity scores	Total missing (N=125)	Minimum	Maximum	Mean	SD	Median	P_{25}	P_{75}	Normality test p-value
APACHE II		4	38	16.34	0.630	15.0	11.0	20.0	<0.001
SAPS II		18	105	49.80	1.244	50.0	41.0	57.0	0.004
SAPS 3		35	100	59.38	1.074	59.0	50.0	66.0	0.049
SOFA		1	17	6.90	0.294	7.0	4.0	9.0	0.005
Shang COVID		1	7	4.24	0.125	4.0	3.0	5.0	<0.001
SEIMC		2	23	9.94	0.450	9.0	6.50	13.0	<0.001
BURDEN	3 (2.4)	1	7	3.17	0.076	3.0	3.0	4.0	<0.001
Inflammation-based	30 (24.0)	1	3	2.49	0.054	3.0	2.0	3.0	<0.001

4.1.3. Hospital mortality

As shown in Table 4.1.3.1, the SEIMC score demonstrates the highest predictive accuracy with an AUC of 0.808 (95% CI: 0.722-0.895, p<0.001). This is followed by the SAPS 3 score, which shows good performance with an AUC of 0.732 (95% CI: 0.630-0.834, p<0.001). The SHANG score performs similarly to SAPS 3, with an AUC of 0.730 (95% CI: 0.628-0.831, p<0.001).The widely used APACHE II score shows moderate predictive ability with an AUC of 0.704 (95% CI: 0.604-0.804, p<0.001), while SAPS II and SOFA scores demonstrate slightly lower performance with AUCs of 0.686 (95% CI: 0.579-0.792, p=0.001) and 0.676 (95% CI: 0.566-0.785, p=0.003), respectively. Notably, the BURDEN score appears to have poor predictive value, with an AUC of 0.505 (95% CI: 0.384-0.625, p=0.939), which is not significantly different from random chance. All scores, except for BURDEN, show statistically significant predictive ability (p<0.05). The narrow confidence intervals for most scores suggest good precision in the AUC estimates.

Table 4.1.3.1 – Estimated AUCs results, p-values and CI for hospital mortality outcome.

Scores	AUC estimate	p-value	95% Confidence Interval	
			Lower Bound	Upper Bound
APACHE II	0.704	<0.001	0.604	0.804
SAPS II	0.686	0.001	0.579	0.792

SAPS 3	0.732	<0.001	0.630	0.834
SOFA	0.676	0.003	0.566	0.785
SEIMC	0.808	<0.001	0.722	0.895
SHANG	0.730	<0.001	0.628	0.831
BURDEN	0.505	0.939	0.384	0.625

Figure 4.1.3.1 represents ROC curves of each score for hospital mortality. The SEIMC score with an AUC of 0.808 indicates good discriminatory power in predicting outcomes. Its ROC curve is positioned closest to the top-left corner of the plot, representing the best balance of sensitivity and specificity among the evaluated scores. SAPS 3 and SHANG scores follow closely behind, with AUCs of 0.732 and 0.730 respectively. Their ROC curves are situated slightly below the SEIMC curve but still demonstrate fair predictive ability. The APACHE II score shows moderate performance with an AUC of 0.704. Its is positioned lower than those of SEIMC, SAPS 3, and SHANG, yet still indicative of better-than-chance prediction. SAPS II and SOFA scores exhibit somewhat lower performance, with AUCs of 0.686 and 0.676 respectively. Their ROC curves are positioned closer to the diagonal reference line compared to the higher-performing scores, suggesting less optimal discrimination between outcomes. In contrast, the BURDEN score shows poor predictive value with an AUC of 0.505. Its ROC curve is very close to the diagonal reference line, indicating performance no better than random chance in predicting outcomes.

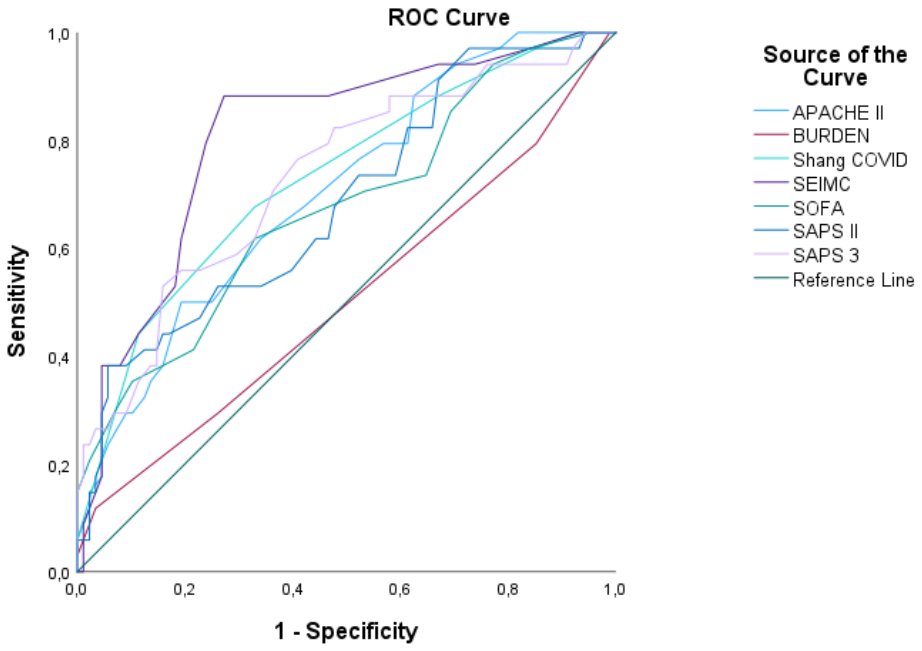

Figure 4.1.3.1 - Severity scores ROC curves for hospital mortality in first COVID-19 wave.

The inflammation-based score was excluded from comparison with the other scores due to a substantially higher number of missing values, which led to a reduced dataset relative to the others. This limitation primarily arose from the absence of albumin values in a significant portion of the cases analysed, as albumin is an essential parameter in the calculation of this score. Consequently, the inflammation-based score could not be reliably assessed alongside the other scoring systems.

Table 4.1.3.2 represents the estimated AUC result, p-value and CI for inflammation-based score in hospital mortality outcome.

Table 4.1.3.2 – Inflammation-based score AUC, p-value and CI for hospital mortality outcome (first wave).

AUC estimate	p-value	95% Confidence Interval	
		Lower Bound	Upper Bound
0.583	0.227	0.452	0.713

Inflammation-based score ROC curve is presented in Figure 4.1.3.2 with an AUC of 0.583, the curve is positioned relatively close to the diagonal line that represents a random guess, implying lack of discriminative ability. A well-performing ROC curve typically arches closer to

the top left corner, showing higher sensitivity and specificity. In this case, however, the curve likely lacks a pronounced bend toward the top left, visually reflecting the low predictive performance and weak classification capability for the score being evaluated.

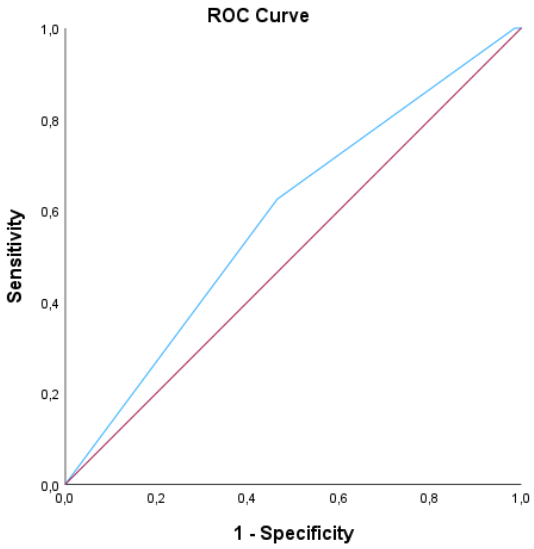

Figure 4.1.3.2 – Inflammation-based score ROC curve for hospital mortality outcome (first wave).

Table 4.1.3.3 presents paired comparisons of the discriminative capacity of various ICU scoring systems: APACHE II, BURDEN, SHANG, SEIMC, SOFA, SAPS II, and SAPS 3. Each cell contains a confidence interval for the difference in discriminative ability between two scores and a p-value indicating the statistical significance of this difference.

APACHE II shows a statistically significant difference in discriminative capacity only when compared to BURDEN (p=0.004). The positive confidence interval (0.065; 0.334) suggests that APACHE II has a higher discriminative ability than BURDEN. Comparisons between APACHE II and other scores do not show statistically significant differences in discriminative capacity.

BURDEN demonstrates statistically significant differences in discriminative ability with all other scores ($p \leq 0.013$ for all comparisons). The negative confidence intervals indicate that BURDEN consistently shows lower discriminative capacity compared to other scores. The most pronounced difference is observed between BURDEN and SEIMC ($p<0.001$).

SHANG does not show statistically significant differences in discriminative ability when compared to other scores, although its comparison with SEIMC approaches significance (p=0.055). SEIMC shows a statistically significant difference in discriminative capacity only with SAPS II (p=0.043). The positive confidence interval (0.004; 0.242) suggests that SEIMC has a slightly higher discriminative ability than SAPS II.

Comparisons among SOFA, SAPS II, and SAPS 3 do not reveal any statistically significant differences in discriminative capacity, indicating that these scores may have similar abilities to distinguish between different patient outcomes or conditions. These findings highlight that while some scores (like SOFA, SAPS II, and SAPS 3) may have comparable discriminative abilities, others (particularly BURDEN) show significant differences.

APACHE II and SEIMC appear to have higher discriminative capacities in certain comparisons.

Table 4.1.3.3 – Paired- test for the null difference in AUCs, to hospital mortality in first COVID-19 wave.

Scores	BURDEN	SHANG	SEIMC	SOFA	SAPS II	SAPS 3
APACHE II	(0.065; 0.334) p=0.004	(-0.140; 0.089) p=0.662	(-0.215; 0.006) p=0.064	(-0.067; 0.124) p=0.557	(-0.050; 0.087) p=0.599	(-0.113; 0.058) p=0.529
BURDEN		(-0.358; -0.092) p=0.001	(-0.433; -0.175) p=<0.001	(-0.306; -0.037) p=0.013	(-0.321; -0.041) p=0.011	(-0.358; -0.097) p=0.001
SHANG			(-0.159; 0.002) p=0.055	(-0.088; 0.197) p=0.456	(-0.070; 0.158) p=0.449	(-0.106; 0.102) p=0.970
SEIMC				(-0.011; 0.277) p=0.071	(0.004; 0.242) p=0.043	(-0.033; 0.186) p=0.170
SOFA					(-0.102; 0.082) p=0.828	(-0.149; 0.037) p=0.236
SAPS II						(-0.131; 0.040) p=0.292

Paired test for the difference in AUCs. Each result corresponds to 95% confidence interval for the AUC difference and corresponding P-value.

4.1.4. ICU mortality

Table 4.1.4.1 presents the scores and their predictive performance based on the estimated Area Under the Curve (AUC), p-value, and 95% Confidence Interval (CI) for each score.

The APACHE II score achieved an AUC of 0.744 (p-value <0.001), and a 95% CI ranging from 0.638 to 0.849. The SAPS II score showed an AUC of 0.718, (p-value =0.001), and a 95% CI from 0.602 to 0.835. For SAPS 3, the AUC was higher at 0.777, (p-value<0.001) and a 95% CI between 0.676 and 0.877.

The SOFA score demonstrated an AUC of 0.729, (p-value <0.001) and a 95% CI spanning 0.619 to 0.838. The SEIMC score had the highest AUC at 0.805, (p-value <0.001) and a 95% CI from 0.719 to 0.891. In comparison, the SHANG score obtained an AUC of 0.703, with a p-value of 0.002 and a 95% CI from 0.590 to 0.815. Lastly, the BURDEN score had the lowest AUC at 0.579, with a p-value of 0.217, and a 95% CI between 0.446 and 0.712, indicating it was not statistically significant in this analysis.

Overall, SEIMC showed the best discriminative ability, while BURDEN performed the least effectively in this outcome.

Table 4.1.4.1 - Estimated AUCs results, p-values and CI for ICU mortality outcome.

Scores	AUC estimate	p-value	95% Confidence Interval	
			Lower Bound	Upper Bound
APACHE II	0.744	<0.001	0.638	0.849
SAPS II	0.718	0.001	0.602	0.835
SAPS 3	0.777	<0.001	0.676	0.877
SOFA	0.729	<0.001	0.619	0.838
SEIMC	0.805	<0.001	0.719	0.891
SHANG	0.703	0.002	0.590	0.815
BURDEN	0.579	0.217	0.446	0.712

Figure 4.1.4.1 presents the ROC curves for ICU mortality for the first wave. The SEIMC score has the highest AUC at 0.805 and its ROC curve indicates strong sensitivity and specificity. Similarly, the SAPS 3 and APACHE II scores have relatively high AUCs of 0.777 and 0.744, respectively, which means their ROC curves also displayed a good degree of separation from the diagonal line, demonstrating notable predictive ability. The SOFA and SAPS II scores have moderate AUCs (0.729 and 0.718, respectively) and their ROC curves are still curve toward the top left, though less prominently than SEIMC, SAPS 3, or APACHE II.

In contrast, the SHANG score, with an AUC of 0.703, and especially the BURDEN score, with an AUC of 0.579, show ROC curves closer to the diagonal, indicating weaker discriminative ability. The BURDEN score has an AUC below 0.6 (p-value = 0.217) and its curve is close to the diagonal line, reflecting a performance close to random chance.

Overall, the visual appearance of the ROC curves for SEIMC, SAPS 3, and APACHE II reflect better classification power with curves that are more convex and oriented toward the top left, while BURDEN's curve is nearly flat, indicating limited effectiveness in distinguishing outcomes.

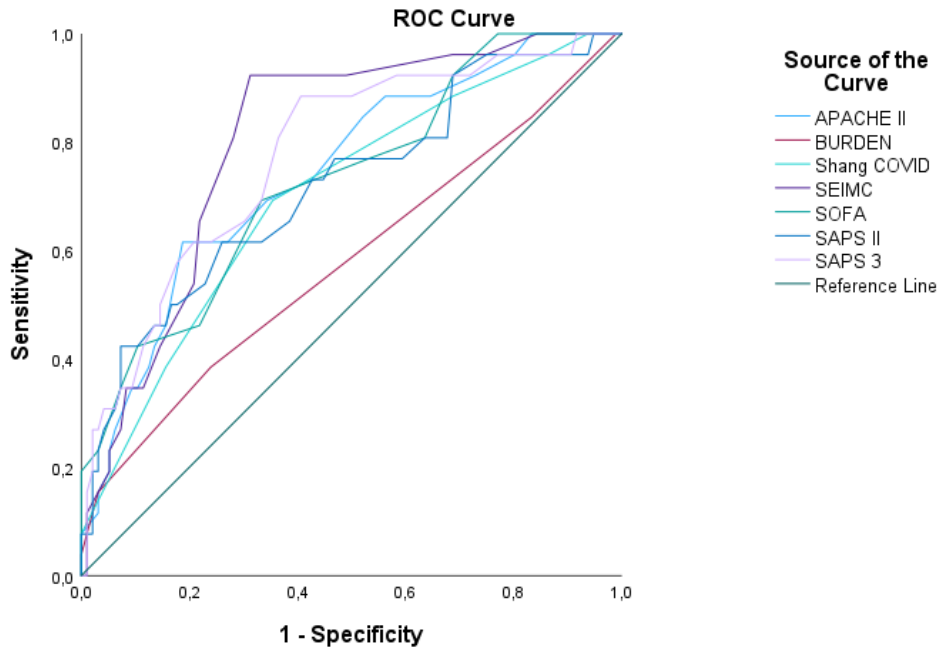

Figure 4.1.4.1 – Severity scores ROC curves for ICU mortality in first wave.

Table 4.1.4.2 represents the estimated AUC result, p-value and CI for inflammation-based score in ICU mortality outcome.

Table 4.1.4.2 - Inflammation-based score estimated AUC, p-value and CI for ICU mortality outcome (first wave).

AUC estimate	p-value	95% Confidence Interval	
		Lower Bound	Upper Bound
0.549	0.512	0.405	0.692

Figure 4.1.4.2 represents inflammation-based score ROC curve (blue line) for ICU mortality outcome. As it can be seen, the line is very close to the reference line (red), meaning lack of discriminative ability.

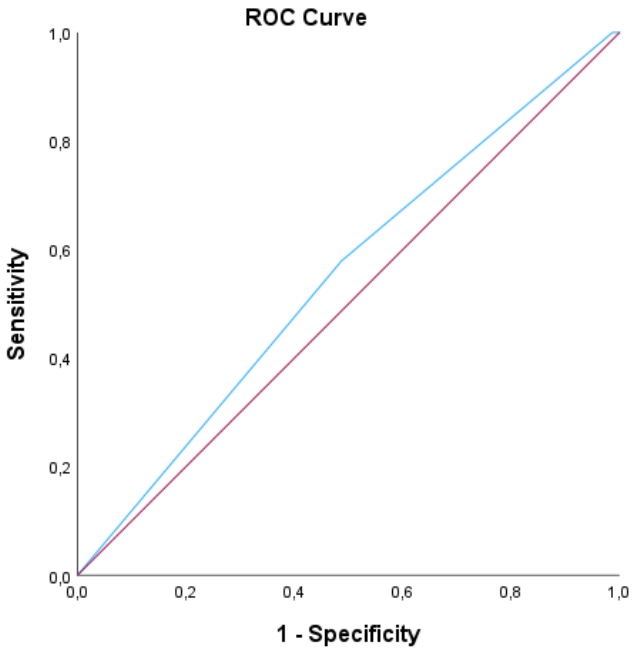

Figure 4.1.4.2 - Inflammation-based ROC curve (blue) for ICU mortality outcome (first wave).

In the Table 4.1.4.3, APACHE II shows a statistically significant difference in discriminative ability compared to BURDEN, with a CI of (0.018; 0.311) and p=0.027. This indicates that APACHE II has a higher discriminative capacity than BURDEN, as the CI does not include zero and is entirely positive.

BURDEN demonstrates significantly lower discriminative capacity compared to several other systems. Its CI when compared to SEIMC is (-0.366; -0.086) with p=0.002, to SOFA is (-0.291; -0.008) with p=0.039, and to SAPS 3 is (-0.342; -0.053) with p=0.007. These negative CIs that do not include zero suggest BURDEN's significant inferior performance.

SHANG shows a significantly lower discriminative capacity compared to SEIMC, with a CI of (-0.194; -0.010) and p=0.030.

Interestingly, the CIs for comparisons among SEIMC, SOFA, SAPS II, and SAPS 3 all include zero and have p-values >0.05, indicating no statistically significant differences in their discriminative capacities. For example, the CI for SEIMC vs. SOFA is (-0.067; 0.220) with p=0.298, and for SAPS II vs. SAPS 3 is (-0.154; 0.037) with p=0.231.

Table 4.1.4.3 - Results for paired-test for the difference in AUCs for ICU mortality outcome (first wave).

Scores	BURDEN	SHANG	SEIMC	SOFA	SAPS II	SAPS 3
APACHE II	(0.018; 0.311) p=0.027	(-0.084; 0.166) p=0.524	(-0.184; 0.061) p=0.328	(-0.070; 0.100) p=0.730	(-0.047; 0.098) p=0.494	(-0.121; 0.055) p=0.459
BURDEN		(-0.265; 0.018) p=0.086	(-0.366; -0.086) p=0.002	(-0.291; -0.008) p=0.039	(-0.298; 0.019) p=0.085	(-0.342; -0.053) p=0.007
SHANG			(-0.194; -0.010) p=0.030	(-0.166; 0.115) p=0.721	(-0.144; 0.113) p=0.814	(-0.177; 0.029) p=0.160
SEIMC				(-0.067; 0.220) p=0.298	(-0.051; 0.225) p=0.219	(-0.093; 0.150) p=0.651
SOFA					(-0.069; 0.090) p=0.801	(-0.136; 0.039) p=0.279
SAPS II						(-0.154; 0.037) p=0.231

Paired test for the difference in AUCs. Each result corresponds to 95% confidence interval for the AUC difference and corresponding P-value.

4.1.5. Early mortality at ICU

Table 4.1.5.1 shows the AUC, p-values, and 95% confidence intervals (CI) for the 7 scores that are comparable. The APACHE II score has an AUC of 0.750, (p-value = 0.003), and a 95% CI ranging from 0.606 to 0.894, indicating good discriminatory ability. Similarly, SAPS II demonstrate an AUC of 0.742, (p-value = 0.004), and a CI between 0.578 and 0.906, suggesting strong predictive performance. SAPS 3 also shows reliable accuracy, with an AUC of 0.738, (p-value = 0.004), and a CI from 0.598 to 0.879.

Among these scores, SOFA demonstrates particularly good predictive ability, with the highest AUC value of 0.760, (p-value = 0.002), and a 95% CI between 0.623 and 0.897. SEIMC achieves the highest overall AUC at 0.786, with a highly significant p-value of 0.001 and a CI ranging from 0.671 to 0.902, indicating excellent discriminatory power. In contrast, SHANG and BURDEN show more limited predictive utility. The SHANG score has an AUC of 0.617 and a non-significant p-value of 0.158, with a 95% CI between 0.451 and 0.783. Meanwhile, BURDEN has the lowest AUC at 0.573, a non-significant p-value of 0.377, and a CI from 0.389 to 0.758, suggesting poor predictive performance.

Table 4.1.5.1 - Estimated AUCs results, p-values and CI for early mortality at ICU outcome (first wave).

Scores	AUC estimate	p-value	95% Confidence Interval	
			Lower Bound	Upper Bound
APACHE II	0.750	0.003	0.606	0.894
SAPS II	0.742	0.004	0.578	0.906
SAPS 3	0.738	0.004	0.598	0.879
SOFA	0.760	0.002	0.623	0.897
SEIMC	0.786	0.001	0.671	0.902
SHANG	0.617	0.158	0.451	0.783
BURDEN	0.573	0.377	0.389	0.758

As shown in Figure 4.1.5.1 for APACHE II, the AUC of 0.750 (p=0.003, CI: 0.606–0.894) shows a reasonable AUC, indicating good discriminative power. SAPS II and SAPS 3 also have relatively high AUC values of 0.742 (p=0.004, CI: 0.578–0.906) and 0.738 (p=0.004, CI: 0.598–0.879), respectively. These scores produce ROC curves (figure 4.2.5) that effectively differentiate between outcomes, with significant p-values supporting the reliability of their AUC values. The SOFA score has an AUC of 0.760, the second highest best result, (p-value = 0.002) and a CI of 0.623–0.897. SEIMC shows the highest AUC at 0.786 (p=0.001, CI: 0.671–0.902), reflecting the most effective ROC curve among these scores and the best excellent discriminative ability.

In contrast, SHANG and BURDEN have weaker ROC curves. SHANG has an AUC of 0.617 (p=0.158, CI: 0.451–0.783), which is below the threshold for good discrimination, and its p-value according indicates limited confidence in its ROC curve's effectiveness. BURDEN has the lowest AUC at 0.573 (p=0.377, CI: 0.389–0.758), suggesting poor discriminative power and an ineffective ROC curve for distinguishing outcomes.

Overall, the ROC curves for SEIMC, SOFA, APACHE II, SAPS II, and SAPS 3 are strong, with SEIMC and SOFA showing the highest AUC values, indicating robust discriminative ability. SHANG and BURDEN, however, have weak ROC curves, with lower AUC values and non-significant p-values, suggesting limited discriminative utility.

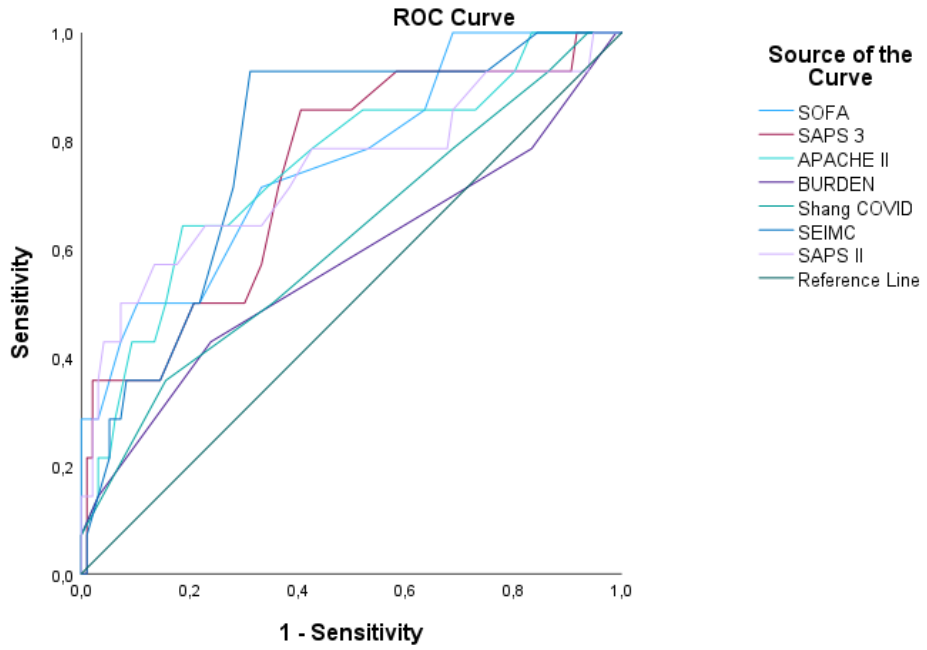

Figure 4.1.5.1 - Severity scores ROC curves for early mortality at ICU outcome (first wave).

Table 4.1.5.2 represents the estimated AUC result, p-value and CI for inflammation-based score in ICU mortality outcome.

Table 4.1.5.2 - Inflammation-based score estimated AUC, p-value and CI for early mortality at ICU outcome (first wave).

AUC estimate	p-value	95% Confidence Interval	
		Lower Bound	Upper Bound
0.537	0.716	0.340	0.735

Figure 4.1.5.2 represents inflammation-based score ROC curve (blue line) for early ICU mortality outcome. As it can be seen, the line is very close to the reference line (red), meaning poor discriminative ability.

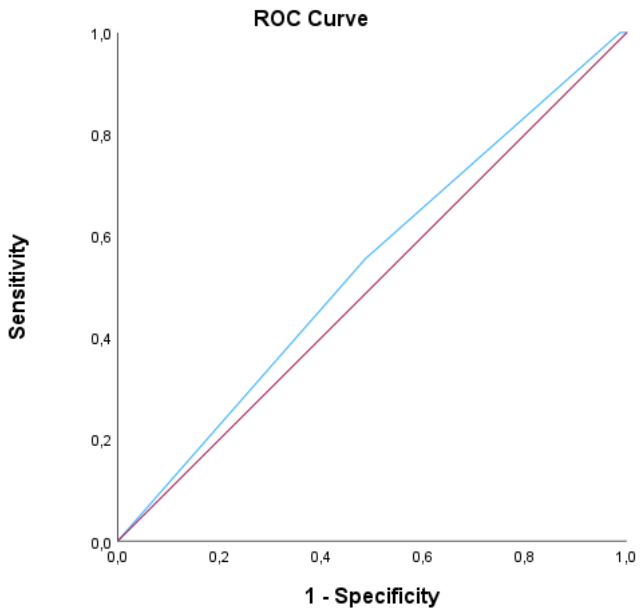

Figure 4.1.5.2 - Inflammation-based score ROC curve for early mortality at ICU outcome (first wave).

In this Table 4.5.1.3, APACHE II demonstrates no statistically significant differences in discriminative ability when compared to any other score. All confidence intervals for APACHE II comparisons include zero, and p-values are consistently above 0.05, suggesting that its performance is comparable to the other scores.

BURDEN, however, shows significantly lower discriminative capacity compared to SEIMC (CI: -0.400; -0.026, p=0.026) and SOFA (CI: -0.364; -0.009, p=0.039). These negative confidence intervals that do not include zero indicate BURDEN's inferior performance relative to these two systems. Comparisons with other systems did not yield statistically significant differences.

SHANG demonstrates a significantly lower discriminative capacity compared to SEIMC (CI: -0.316; -0.022, p=0.025). This suggests that SHANG may be less effective than SEIMC in distinguishing between different patient outcomes or conditions in the ICU. However, SHANG's performance appears comparable to the other scores in the study.

Notably, SEIMC, SAPS II, SAPS 3, and SOFA do not show statistically significant differences in discriminative capacity when compared to each other. All confidence intervals for these comparisons include zero, and p-values are well above the conventional 0.05 threshold. This suggests that these four scoring systems may have similar abilities to distinguish between different patient outcomes or conditions in intensive care units.

While some differences exist, particularly involving BURDEN and SHANG, many of the commonly used systems (APACHE II, SEIMC, SAPS II, SAPS 3, and SOFA) appear to have comparable discriminative capacities in this analysis.

Table 4.1.5.3 - Results for paired-test for the difference in AUCs for early mortality at ICU outcome (first wave).

Scores	BURDEN	SHANG	SEIMC	SAPS II	SAPS 3	SOFA
APACHE II	(-0.015; 0.368) p=0.070	(-0.023; 0.288) p=0.094	(-0.209; 0.137) p=0.682	(-0.100; 0.116) p=0.882	(-0.075; 0.099) p=0.788	(-0.096; 0.076) p=0.820
BURDEN		(-0.231; 0.144) p=0.646	(-0.400; -0.026) p=0.026	(-0.375; 0.038) p=0.110	(-0.370; 0.041) p=0.116	(-0.364; -0.009) p=0.039
SHANG			(-0.316; -0.022) p=0.025	(-0.274; 0.025) p=0.103	(-0.254; 0.012) p=0.076	(-0.291; 0.005) p=0.059
SEIMC				(-0.152; 0.241) p=0.659	(-0.128; 0.224) p=0.593	(-0.152; 0.204) p=0.774
SAPS II					(-0.108; 0.115) p=0.948	(-0.108; 0.071) p=0.689
SAPS 3						(-0.115; 0.071) p=0.643

4.1.6. Late mortality at ICU

Table 4.1.6.1 presents the estimated AUCs, p-value and CI for late mortality at ICU outcome (first wave). APACHE II has an AUC of 0.736 (p-value= 0.008) and a CI from 0.601 to 0.872, suggesting moderate discriminative ability.

SAPS II shows a slightly lower AUC of 0.691, also statistically significant (p=0.032), with a CI between 0.543 and 0.839, indicating fair accuracy.

SAPS 3 has a high AUC of 0.822, with a highly significant p-value of less than 0.001 and a CI from 0.701 to 0.943, reflecting strong discriminative power. Similarly, SEIMC achieves the highest AUC at 0.827, with a significant p-value of less than 0.001 and a CI between 0.727 and 0.927, indicating excellent predictive performance.

The SOFA score has an AUC of 0.692, with a p-value of 0.031 and a CI from 0.534 to 0.850, suggesting fair discriminative ability. SHANG also performs well, with an AUC of 0.803, a significant p-value of 0.001, and a CI from 0.702 to 0.904, indicating strong predictive accuracy.

In contrast, BURDEN has a low AUC of 0.586 with a non-significant p-value of 0.333 and a CI between 0.410 and 0.762, suggesting weak discriminative ability and limited utility in outcome prediction.

Overall, SEIMC, SAPS 3, and SHANG show the strongest ROC curves, with high AUC values and significant p-values, indicating effective discriminative power. APACHE II, SAPS II, and SOFA demonstrate moderate accuracy, while BURDEN has the weakest performance, with a low AUC and non-significant p-value.

Table 4.1.6.1 - Estimated AUC, p-value and CI for late mortality at ICU outcome (first wave).

Scores	AUC estimate	p-value	95% Confidence Interval	
			Lower Bound	Upper Bound
APACHE II	0.736	0.008	0.601	0.872
SAPS II	0.691	0.032	0.543	0.839
SAPS 3	0.822	<0.001	0.701	0.943
SOFA	0.692	0.031	0.534	0.850
SEIMC	0.827	<0.001	0.727	0.927
SHANG	0.803	0.001	0.702	0.904
BURDEN	0.586	0.333	0.410	0.762

Figure 4.1.6.1 shows the severity scores ROC curves for late mortality at ICU outcome (first wave). SEIMC demonstrates the strongest ROC curve, with the highest AUC of 0.827 (p-value < 0.001). Its confidence interval (0.727 to 0.927) further supports its excellent discriminative power, indicating that the SEIMC score is highly reliable for outcome prediction.

SAPS 3 also shows a robust ROC curve with an AUC of 0.822, (p-value < 0.001), and a confidence interval of 0.701 to 0.943, suggesting strong discriminatory ability and reliable predictive accuracy.

SHANG presents a similarly strong ROC curve, with an AUC of 0.803, a significant p-value of 0.001, and a CI between 0.702 and 0.904. This reflects good predictive accuracy, though slightly lower than SEIMC and SAPS 3.

APACHE II has a moderate AUC of 0.736, with a p-value of 0.008 and a CI of 0.601 to 0.872, indicating reasonable discriminative ability but not as strong as the top-performing scores.

SAPS II and SOFA both show fair ROC curve performance, with AUCs of 0.691 and 0.692, respectively. SAPS II has a p-value of 0.032, and SOFA has a p-value of 0.031, with confidence intervals (SAPS II: 0.543 to 0.839; SOFA: 0.534 to 0.850) that indicate moderate discriminative capability.

In contrast, BURDEN has the weakest ROC curve, with an AUC of 0.586, a non-significant p-value of 0.333, and a confidence interval of 0.410 to 0.762. This low AUC suggests that BURDEN has poor discriminative power and is ineffective for reliable outcome prediction.

In summary, SEIMC, SAPS 3, and SHANG demonstrate the strongest ROC curves with high AUC values and significant p-values, indicating strong discriminative power. APACHE II, SAPS II, and SOFA show moderate ROC curve performance, while BURDEN has the weakest ROC curve, with limited utility for outcome discrimination.

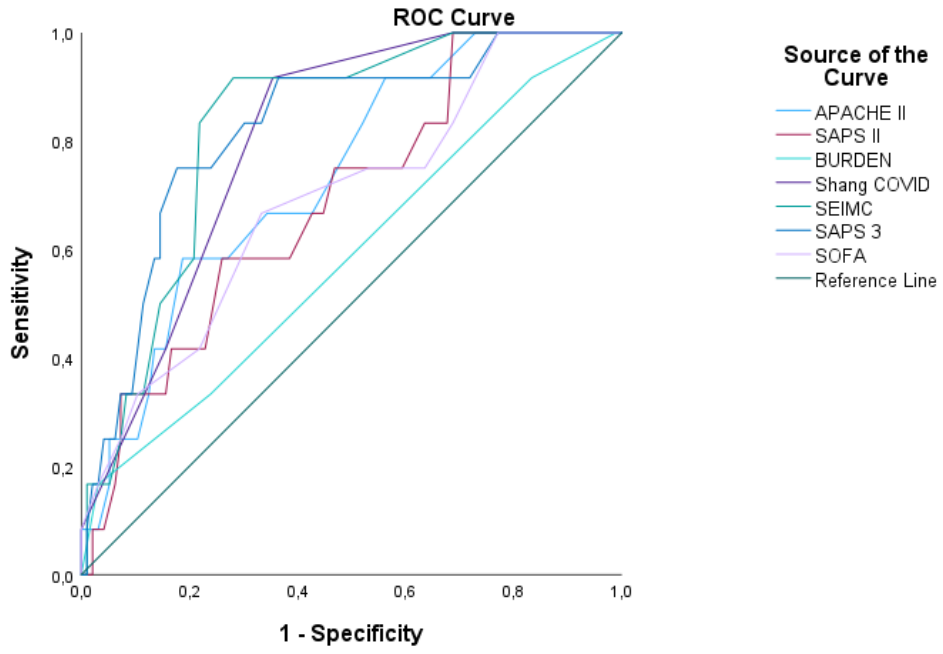

Figure 4.1.6.1 - Severity scores ROC curves for late mortality at ICU outcome (first wave).

Table 4.1.6.2 represents the estimated AUC result, p-value and CI for inflammation-based score in late mortality at ICU outcome (first wave).

Table 4.1.6.2 - Inflammation-based score estimated AUC, p-value and CI for late mortality at ICU outcome (first wave).

AUC estimate	p-value	95% Confidence Interval	
		Lower Bound	Upper Bound
0.559	0.544	0.372	0.746

Figure 4.1.6.2 represents inflammation-based score ROC curve (blue line) for late mortality at ICU outcome. As it can be seen, the line is very close to the reference line (red), meaning lack of discriminative ability.

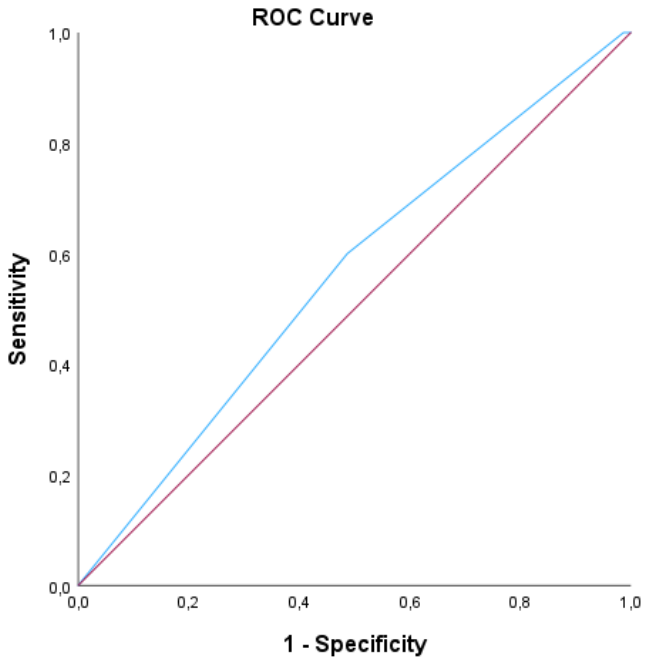

Figure 4.1.6.2 - Inflammation-based score ROC curve for late mortality at ICU outcome (first wave).

In Table 4.1.6.3, APACHE II and SAPS II do not show statistically significant differences in discriminative ability when compared to any other scoring system. All p-values for their comparisons are above 0.05, and all CIs include zero, suggesting comparable performance across systems.

BURDEN demonstrates significantly lower discriminative capacity compared to SHANG (CI: -0.403; -0.031, p=0.022), SAPS 3 (CI: -0.432; -0.050, p=0.013), and SEIMC (CI: -0.423; -0.049, p=0.013). The negative CIs that do not include zero suggest BURDEN's inferior performance relative to these three systems.

SHANG shows no statistically significant differences in discriminative capacity when compared to other systems, except for the comparison with BURDEN. SAPS 3 shows a significantly higher discriminative capacity compared to SOFA (CI: 0.007; 0.254, p=0.039). This positive CI that doesn't include zero indicates SAPS 3's superior performance relative to SOFA. SOFA and SEIMC do not show statistically significant differences in discriminative capacity when compared to other systems, except for the comparisons mentioned above.

Table 4.1.6.3 - Results for paired-test for the difference in AUCs for late mortality at ICU outcome (first wave).

Scores	SAPS II	BURDEN	SHANG	SAPS 3	SOFA	SEIMC
APACHE II	(-0.032; 0.122) p=0.249	(-0.058; 0.358) p=0.157	(-0.237; 0.104) p=0.442	(-0.233; 0.061) p=0.251	(-0.086; 0.175) p=0.506	(-0.254; 0.072) p=0.275
SAPS II		(-0.131; 0.341) p=0.383	(-0.297; 0.073) p=0.236	(-0.276; 0.014) p=0.077	(-0.114; 0.113) p=0.988	(-0.319; 0.047) p=0.146
BURDEN			(-0.403; -0.031) p=0.022	(-0.432; -0.050) p=0.013	(-0.320; 0.108) p=0.332	(-0.423; -0.049) p=0.013
SHANG				(-0.167; 0.128) p=0.800	(-0.101; 0.323) p=0.304	(-0.086; 0.039) p=0.454
SAPS 3					(0.007; 0.254) p=0.039	(-0.163; 0.154) p=0.953
SOFA						(-0.346; 0.076) p=0.210

Paired test for the difference in AUCs. Each result corresponds to 95% confidence interval for the AUC difference and corresponding P-value.

4.1.7. Discussion of the results within first wave

In a summary way and to facilitate the comparison of the discriminative performance of each score for the different outcomes, a summary of the results is presented in Table 4.1.7.1.

The APACHE II score demonstrated the highest area under the curve (AUC) for ICU mortality (0.744), with the AUC for hospital mortality being slightly lower at 0.704. It showed relatively consistent predictive performance for early (0.750) and late ICU mortality (0.736). The SAPS II score, while slightly less accurate than APACHE II for hospital mortality (0.686), exhibited similar results for ICU mortality (0.718) and early ICU mortality (0.742), with a somewhat lower AUC for late ICU mortality (0.691).

The SAPS 3 score performed well across all mortality categories, with the highest AUC for late ICU mortality (0.822) and ICU mortality (0.777). However, it had a slightly lower AUC for early ICU mortality (0.738) compared to the other scores.

SOFA, on the other hand, had lower AUCs across most categories, with the highest value being for early ICU mortality (0.760), but it had a lower AUC for late ICU mortality (0.692). The SOFA is used to assess organ dysfunction, which makes it more relevant for early mortality since organ failure typically occurs in the early stages of severe illness, before a fatal outcome. This score is often used to monitor critically ill patients, such as those in intensive care units. When used in situations with an immediate risk of death, the SOFA score can quickly identify clinical deterioration and provide guidance for urgent treatment. For this reason, SOFA is particularly effective in predicting early mortality, being an important indicator in the early phases of the disease when prompt intervention may be critical to saving lives.

The SEIMC score showed the strongest predictive ability overall, with the highest AUC for both hospital mortality (0.808) and ICU mortality (0.805), along with a high AUC for late ICU mortality (0.827). It was also predictive of early ICU mortality (0.786). Finally, the Shang score had moderate performance across the categories, with the highest AUC for late ICU mortality (0.803) and the lowest AUC for early ICU mortality (0.617).

In summary, SEIMC generally outperforms the other scoring systems across all mortality categories, while SAPS 3 performs well for ICU mortality and late ICU mortality. APACHE II and SAPS II show moderate predictive ability, and SOFA and Shang scores perform less robustly across the different mortality outcomes.

Table 4.1.7.1 – Estimated AUCs and confidence intervals results for mortality outcomes, in first wave.

Scores	Hospital Mortality	ICU Mortality	Early ICU Mortality	Late ICU Mortality
APACHE II	0.704 (0.604-0.804)	0.744 (0.638-0.849)	0.750 (0.606-0.894)	0.736 (0.601-0.872)
SAPS II	0.686 (0.579-0.792)	0.718 (0.602-0.835)	0.742 (0.578-0.906)	0.691 (0.543-0.839)
SAPS 3	0.732 (0.630-0.834)	0.777 (0.676-0.877)	0.738 (0.598-0.879)	0.822 (0.701-0.943)
SOFA	0.676 (0.566-0.785)	0.729 (0.619-0.838)	0.760 (0.623-0.897)	0.692 (0.534-0.850)
SEIMC	0.808 (0.722-0.895)	0.805 (0.719-0.891)	0.786 (0.671-0.902)	0.827 (0.727-0.927)
SHANG	0.730 (0.628-0.831)	0.703 (0.590-0.815)	0.617 (0.451-0.783)	0.803 (0.702-0.904)

4.2. Results for Second Wave Data

4.2.1. Clinical and demographic characteristics

Table 4.2.1.1 presents a comprehensive overview of patient characteristics, outcomes, and severity scores for a cohort of 161 COVID-19 patients admitted to the intensive care unit (ICU). The median age of the patients is 67 years (P_{25} = 55.0; P_{75}= 76.5). The distribution shows a male predominance, with 115 (71.4%) of the patients being male and 46 (28.6%) female.

A large majority of patients, 138 (85.7%), had at least one comorbidity, while 23 (14.3%) had no reported comorbidities. Regarding respiratory support, 95 (59%) patients required invasive mechanical ventilation (IMV), and 20 (12.4%) needed extracorporeal membrane oxygenation (ECMO), indicating the severity of their condition.

The outcomes data reveal a high mortality rate, with 64 (39.8%) patients dying in the ICU. Of these, 27 (16.8%) were classified as early deaths and 37 (23.0%) as late deaths. The total hospital mortality includes 77 patients (47.8%). The median length of stay in the ICU was 8 days (P_{25} = 4.0; P_{75}= 14.0).

The APACHE II score (with 3 missing values, 1.9% of the cohort) had a median of 15 (P_{25} = 11.0; P_{75}= 19.25). The SAPS II score (also with 3 missing values) showed a median of 48 (P_{25} = 37.0; P_{75}= 58.0), while the SAPS 3 score (1 missing value, 0.6%) had a median of 59.50 (P_{25} = 52.0; P_{75}= 68.0). The median SOFA score was 6 (P_{25} = 3.0; P_{75}= 8.50). Shang-COVID score (2 missing values, 1.2%) had a median of 4 (P_{25} = 3.0; P_{75}= 5.0), and the SEIMC score's median was 9 (P_{25} = 6.0; P_{75}= 13.0). The BURDEN score (2 missing values, 1.2%) showed a median of 3 (P_{25} = 2.0; P_{75}= 4.0). The Inflammation-based score had the highest number of missing values (39, representing 24.2% of the cohort) and a median of 2 (P_{25} = 2.0; P_{75}= 3.0).

Table 4.2.1.1 - Clinical characteristics and demographics of second COVID-19 wave.

Variables	Total Missings (%)	All Patients (N=161)
Age, years		67.0 (55.0 - 76.50)
Gender		
Male		115 (71.4)
Female		46 (28.6)
Presence of Comorbidities		
Yes		138 (85.7)
No		23 (14.3)
Respiratory support		
IMV		95 (59.0)
ECMO		20 (12.4)

Outcomes		
Total deceased at ICU		64 (39.8)
Early deceased at ICU		27 (16.8)
Late deceased at ICU		37 (23.0)
Total deceased at hospital		77 (47.8)
Days in ICU		8.0 (4.0-14.0)
Severity scores		
APACHE II	3 (1.9)	15.0 (11.0 - 19.25)
SAPS II	3 (1.9)	48.0 (37.0 - 58.0)
SAPS 3	1 (0.6)	59.50 (52.0 - 68.0)
SOFA		6.0 (3.0 - 8.50)
Shang-COVID	2 (1.2)	4.00 (3.00 - 5.00)
SEIMC		9.0 (6.0 - 13.0)
BURDEN	2 (1.2)	3.0 (2.0 - 4.0)
Inflammation-based	39 (24.2)	2.0 (2.0 - 3.0)

4.2.2. Statistical description for the severity scores for second wave

Table 4.2.2.1 presents a comprehensive summary of various severity scores used in critical care settings for a cohort of 161 patients. These scores are commonly used to assess the severity of illness and predict outcomes in critically ill patients. The APACHE II (Acute Physiology and Chronic Health Evaluation II) score, with 3 missing values (1.9% of the cohort), ranges from 2 to 33. It has a mean of 15.68 (standard deviation 0.501) and a median of 15.0 (P_{25} = 11.0; P_{75}= 19.25. This suggests a moderate severity of illness in the cohort on average. The SAPS II (Simplified Acute Physiology Score II), also with 3 missing values, shows a wider range from 13 to 90. Its mean is 48.33 (SD 1.183) with a median of 48.0 (P_{25} = 37.0; P_{75}= 58.0). This score indicates a relatively high severity of illness in the patient group. SAPS 3 has only 1 missing value (0.6%) and ranges from 35 to 103. Its mean is 60.63 (SD 1.035) with a median of 59.50 (P_{25} = 52.0; P_{75}= 68.0).

The SOFA (Sequential Organ Failure Assessment) score has no missing values and ranges from 0 to 15. Its mean is 6.06 (SD 0.268) with a median of 6.0 (P_{25} = 3.0; P_{75}= 8.5). This indicates moderate organ dysfunction across the cohort.

The Shang COVID score, specifically designed for COVID-19 patients, has 2 missing values (1.2%) and ranges from 1 to 7. Its mean is 4.25 (SD 0.096) with a median of 4.0 (P_{25} = 3.0; P_{75}= 5.0). This suggests moderate severity in the context of COVID-19.

The SEIMC score has no missing values and ranges from 2 to 26. Its mean is 9.89 (SD 0.420) with a median of 9.0 (P_{25} = 6.0; P_{75}= 13.0). This indicates moderate to high severity according to this scoring system. The BURDEN score, with 2 missing values, ranges from 1 to

6. Its mean is 3.18 (SD 0.093) with a median of 3.0 (P_{25} = 2.0; P_{75}= 4.0). This suggests moderate severity based on this scoring system. The Inflammation-based score has the highest number of missing values at 39 (24.2% of the cohort), which could potentially affect its reliability. It ranges from 0 to 3, with a mean of 2.34 (SD 0.062) and a median of 2.0 (P_{25} = 2.0; P_{75}= 3.0), indicating a high level of inflammation in most patients.

Importantly, normality tests indicate that all scores except SAPS II (p=0.074) deviate significantly from a normal distribution (p<0.001 for all others, p=0.005 for SAPS 3). This suggests that non-parametric statistical methods may be more appropriate for analysing these scores in further analyses. The non-normal distribution of most scores also implies that median and IQR values may be more representative of the central tendency and spread of these scores than mean and standard deviation.

Table 4.2.2.1 – Statistical description of second wave data.

Severity scores	Total missings Count (%) (N=161)	Minimum	Maximum	Mean	SD	Median	P_{25}	P_{75}	Normality test p-value
APACHE II	3 (1.9)	2	33	15.68	0.501	15.0	11.0	19.25	<0.001
SAPS II	3 (1.9)	13	90	48.33	1.183	48.0	37.0	58.0	0.074
SAPS 3	1 (0.6)	35	103	60.63	1.035	59.50	52.0	68.0	0.005
SOFA		0	15	6.06	0.268	6.0	3.0	8.50	<0.001
Shang COVID	2 (1.2)	1	7	4.25	0.096	4.0	3.0	5.0	<0.001
SEIMC		2	26	9.89	0.420	9.0	6.0	13.0	<0.001
BURDEN	2 (1.2)	1	6	3.18	0.093	3.0	2.0	4.0	<0.001
Inflammation-based	39 (24.2)	0	3	2.34	0.062	2.0	2.0	3.0	<0.001

4.2.3. Hospital mortality

As presented in Table 4.2.3.1, the p-value represents the probability that the observed AUC value could have occurred by chance, assuming that there is no actual difference in the ability of the scores to discriminate between outcomes. A p-value < 0.05 typically indicates that the observed AUC is statistically significant, meaning the score is likely to have genuine discriminatory ability.

APACHE II has an AUC of 0.769 and a p-value of < 0.001, which is statistically significant. This suggests that APACHE II has a strong ability to discriminate between outcomes, and this result is unlikely to have occurred by chance.

SAPS II shows an AUC of 0.755 and a p-value of < 0.001, which is also statistically significant. This indicates that SAPS II effectively discriminates between outcomes, and the result is unlikely to have occurred by chance.

SAPS 3 has an AUC of 0.759 and a p-value of < 0.001, making it statistically significant as well. The result confirms that SAPS 3 has a good discriminatory ability in predicting outcomes, with a very low likelihood that the observed AUC occurred by random chance.

SOFA has an AUC of 0.717 and a p-value of < 0.001, indicating statistical significance. This suggests that SOFA is effective in discriminating between outcomes, but with a somewhat lower discriminatory ability compared to APACHE II, SAPS II, and SAPS 3.

SEIMC has an AUC of 0.724 and a p-value of < 0.001, which is statistically significant. This indicates that SEIMC also demonstrates good discriminatory power, though still weaker than APACHE II, SAPS II, and SAPS 3.

SHANG shows an AUC of 0.648 and a p-value of 0.001, which is statistically significant. While SHANG is still able to discriminate between outcomes, its discriminatory ability is weaker compared to the other scores with higher AUC values.

BURDEN, however, has an AUC of 0.564 and a p-value of 0.165, which is not statistically significant. This means that the BURDEN score does not demonstrate a statistically significant ability to discriminate between outcomes.

Table 4.2.3.1 - Estimated AUCs, p-value and CI for hospital mortality outcome (second wave).

Scores	AUC estimate	p-value	95% Confidence Interval	
			Lower Bound	Upper Bound
APACHE II	0.769	<0.001	0.695	0.843
SAPS II	0.755	<0.001	0.680	0.830
SAPS 3	0.759	<0.001	0.683	0.834
SOFA	0.717	<0.001	0.638	0.797
SEIMC	0.724	<0.001	0.642	0.805
SHANG	0.648	0.001	0.561	0.734
BURDEN	0.564	0.165	0.474	0.654

Figure 4.2.3.1 shows that APACHE II achieves the highest AUC of 0.769, suggesting it has the strongest overall predictive accuracy among the scores. The p-value (<0.001) and the confidence interval ranging from 0.695 to 0.843 imply that the difference in AUC compared

to the other scores is statistically significant. SAPS II and SAPS 3 show similar AUCs of 0.755 and 0.759, respectively, with p-values of <0.001, indicating their AUCs are significantly different from lower-performing scores. The confidence intervals (0.680–0.830 for SAPS II and 0.683–0.834 for SAPS 3) are comparable to APACHE II, suggesting a similar level of reliable predictive ability.

The SOFA and SEIMC scores have slightly lower AUCs of 0.717 and 0.724, respectively, with p-values of <0.001, showing their AUCs are significantly different from weaker scores. Their confidence intervals (0.638–0.797 for SOFA and 0.642–0.805 for SEIMC) are slightly narrower, suggesting moderate but consistent predictive ability.

The SHANG score, with an AUC of 0.648 and a p-value of 0.001, indicates a weaker predictive power compared to the higher-performing scores. Its confidence interval (0.561–0.734) is wider, reflecting more variability and less reliable performance.

Finally, the BURDEN score has the lowest AUC at 0.564, with a p-value of 0.165, indicating no significant difference from a score with no predictive ability (AUC of 0.5). Its confidence interval of 0.474 to 0.654 crosses below 0.5, suggesting its predictive accuracy is not sufficient for practical use.

In summary, APACHE II, SAPS II, and SAPS 3 show the most reliable predictive accuracy, with significant differences in AUC compared to the weaker scores. SOFA and SEIMC provide moderate predictive value, while SHANG and BURDEN show limited predictive accuracy, with BURDEN demonstrating the least potential for reliable prediction

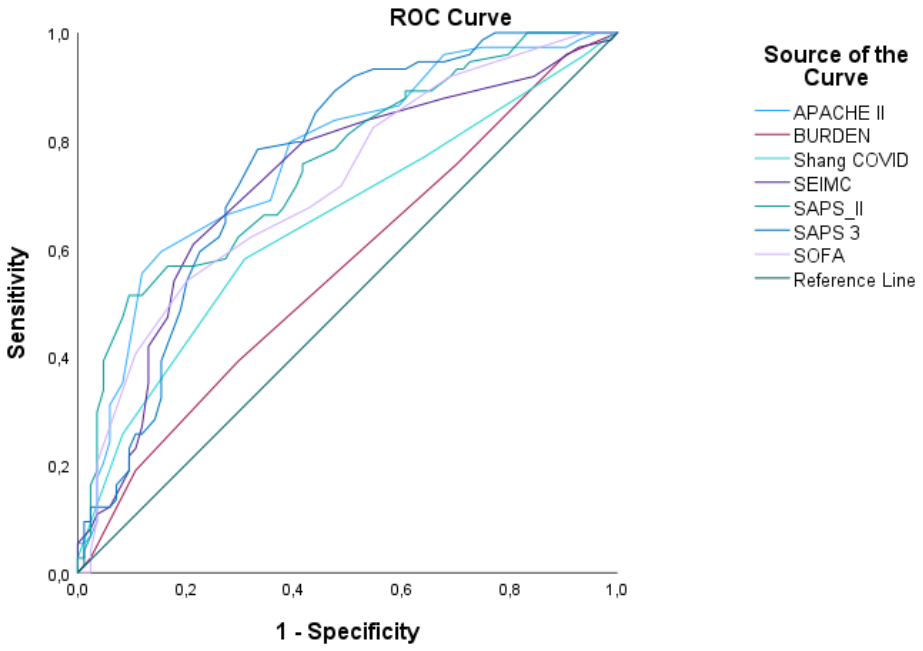

Figure 4.2.3.1 - Severity scores ROC curves for hospital mortality outcome (second wave).

Table 4.2.3.2 represents the estimated AUC result, p-value and CI for inflammation-based score in hospital mortality outcome (second wave).

Table 4.2.3.2 - Inflammation-based score estimated AUC, p-value and CI for hospital mortality outcome (second wave).

AUC estimate	p-value	95% Confidence Interval	
		Lower Bound	Upper Bound
0.483	0.749	0.380	0.586

Figure 4.2.3.2 represents inflammation-based score ROC curve (blue line) for hospital mortality outcome (second wave). As it can be seen, the line is very close to the reference line (red), meaning lack of discriminative ability.

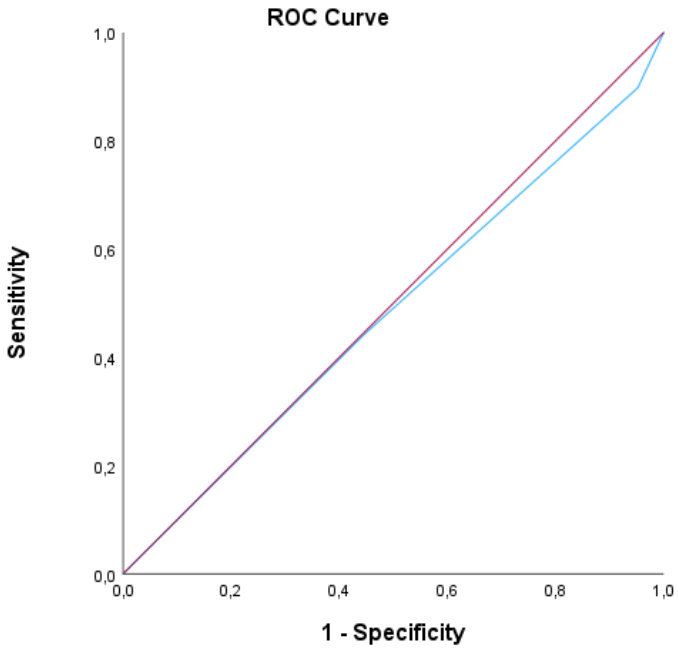

Figure 4.2.3.2 - Inflammation-based score ROC curve for hospital mortality outcome (second wave).

Table 4.2.3.3 shows that APACHE II demonstrates significantly higher discriminative ability compared to BURDEN (CI: 0.089; 0.321, p=0.001) and SHANG (CI: 0.023; 0.220, p=0.016), indicating its superior performance in distinguishing between patient outcomes or conditions relative to these two systems. However, APACHE II shows no statistically significant differences when compared to SEIMC, SAPS II, SAPS 3, or SOFA, suggesting comparable performance with these scores.

BURDEN consistently exhibits lower discriminative capacity across multiple comparisons. It shows significantly inferior performance compared to SEIMC (CI: -0.257; -0.062, p=0.001), SAPS II (CI: -0.303; -0.078, p=0.001), SAPS 3 (CI: -0.300; -0.089, p<0.001), and SOFA (CI: -0.264; -0.042, p=0.007). This consistent pattern suggests that BURDEN may be less effective in differentiating between patient outcomes in intensive care settings compared to other scoring systems.

SHANG also demonstrates some limitations in its discriminative capacity, showing significantly lower performance compared to SAPS II (CI: -0.202; -0.012, p=0.027) and SAPS 3 (CI: -0.202; -0.020, p=0.017). Its comparison with SEIMC approaches statistical significance (CI: -0.152; <0.001, p=0.051), further suggesting potential limitations in its discriminative ability.

Notably, SEIMC, SAPS II, SAPS 3, and SOFA do not show statistically significant differences in discriminative capacity when compared to each other. All confidence intervals for these comparisons include zero, and p-values are consistently above the conventional 0.05

threshold. This suggests that these four scoring systems may have similar abilities to distinguish between different patient outcomes or conditions in intensive care units.

Table 4.2.3.3 - Results for paired-test for the difference in AUCs for hospital mortality outcome (second wave).

Scores	BURDEN	SHANG	SEIMC	SAPS II	SAPS 3	SOFA
APACHE II	(0.089; 0.321) p=0.001	(0.023; 0.220) p=0.016	(-0.049; 0.140) p=0.344	(-0.052; 0.081) p=0.671	(-0.071; 0.092) p=0.800	(-0.034; 0.137) p=0.238
BURDEN		(-0.196; 0.029) p=0.146	(-0.257; -0.062) p=0.001	(-0.303; -0.078) p=0.001	(-0.300; -0.089) p=<0.001	(-0.264; -0.042) p=0.007
SHANG			(-0.152; <0.001) p=0.051	(-0.202; -0.012) p=0.027	(-0.202; -0.020) p=0.017	(-0.178; 0.038) p=0.205
SEIMC				(-0.122; 0.060) p=0.507	(-0.120; 0.050) p=0.421	(-0.097; 0.109) p=0.906
SAPS II					(-0.081; 0.073) p=0.919	(-0.034; 0.108) p=0.307
SAPS 3						(-0.037; 0.119) p=0.303

Paired test for the difference in AUCs. Each result corresponds to 95% confidence interval for the AUC difference and corresponding P-value

4.2.4. ICU mortality

Table 4.2.4.1 presents the Area Under the Curve (AUC) values for several severity scores, illustrating each score's performance in distinguishing outcomes. The APACHE II score has the highest AUC at 0.742, with a p-value <0.001 and a confidence interval of 0.663 to 0.821, suggesting it offers the strongest predictive accuracy among the scores. SAPS II follows closely, with an AUC of 0.700 (p <0.001, 95% CI: 0.617–0.783), indicating its robust predictive capability as well.

SAPS 3 shows a similar AUC of 0.713 (p <0.001, 95% CI: 0.634–0.792), which aligns closely with SAPS II and APACHE II in predictive ability. The SOFA score, with an AUC of 0.692 (p <0.001, 95% CI: 0.607–0.777), also provides a moderate level of predictive accuracy, though slightly lower than the previous scores. SEIMC, with an AUC of 0.706 (p <0.001, 95% CI: 0.622–0.790), demonstrates comparable predictive capacity to SAPS II and SAPS 3.

The SHANG score, with an AUC of 0.642 (p=0.003, 95% CI: 0.553–0.731), reflects a more limited predictive power compared to the other scores. Lastly, the BURDEN score shows

the lowest AUC at 0.602 (p=0.031, 95% CI: 0.513–0.691), indicating that it has the least predictive accuracy among the evaluated scores.

In summary, APACHE II, SAPS II, SAPS 3, and SEIMC demonstrate higher AUC values, suggesting stronger discriminatory power, while SOFA and SHANG show moderate accuracy. The BURDEN score ranks lowest in predictive performance in this assessment.

Table 4.2.4.1 - Estimated AUCs, p-value and CI for ICU mortality outcome (second wave).

Scores	AUC estimate	p-value	95% Confidence Interval	
			Lower Bound	Upper Bound
APACHE II	0.742	<0.001	0.663	0.821
SAPS II	0.700	<0.001	0.617	0.783
SAPS 3	0.713	<0.001	0.634	0.792
SOFA	0.692	<0.001	0.607	0.777
SEIMC	0.706	<0.001	0.622	0.790
SHANG	0.642	0.003	0.553	0.731
BURDEN	0.602	0.031	0.513	0.691

Figure 4.2.4.1 presents the ROC curves for each severity score and provide a visual and statistical assessment of predictive accuracy for ICU mortality in the second wave. The area under the ROC curve (AUC) is the primary metric used here, as it reflects each score's ability to distinguish between survivors and non-survivors.

The ROC curve for APACHE II, which achieves the highest AUC at 0.742, illustrates a strong balance between sensitivity (true positive rate) and specificity (true negative rate), showing that it consistently ranks patients with a higher likelihood of mortality accurately. This high AUC, accompanied by a narrower confidence interval (0.663 to 0.821), suggests that the APACHE II score has a strong and reliable predictive performance.

The SAPS 3 and SEIMC scores show slightly lower AUC values (0.713 and 0.706, respectively), which is reflected in their ROC curves as somewhat less separation between true positive and false positive rates. Their ROC curves are still well above the diagonal line, which represents random chance (AUC = 0.5), indicating that both scores still provide meaningful predictions for ICU mortality but with a lower degree of accuracy than APACHE II.

SAPS II and SOFA scores, with AUCs of 0.700 and 0.692, also have ROC curves that demonstrate moderate predictive accuracy. While not as precise as APACHE II, these curves indicate that SAPS II and SOFA are useful for predicting ICU mortality but may be slightly less

effective in capturing the true sensitivity-specificity trade-off compared to the top-performing scores.

In contrast, the ROC curves for SHANG and BURDEN exhibit notably lower AUCs (0.642 and 0.602, respectively), which reflect a weaker predictive power. The SHANG curve indicates moderate predictive ability, but with more variability, suggesting its predictions are less consistent. BURDEN's ROC curve, which is close to the 0.5 line, suggests it has limited predictive accuracy, performing only marginally better than random classification.

Overall, the ROC curves emphasize that APACHE II provides the most reliable and consistent predictive accuracy for ICU mortality, followed by SAPS 3, SEIMC, SAPS II, and SOFA. SHANG and BURDEN exhibit weaker predictive capabilities, as shown by their lower AUCs and closer proximity to the random chance line on their ROC curves. This highlights the varying effectiveness of these scoring systems for predicting ICU mortality outcomes in the second wave.

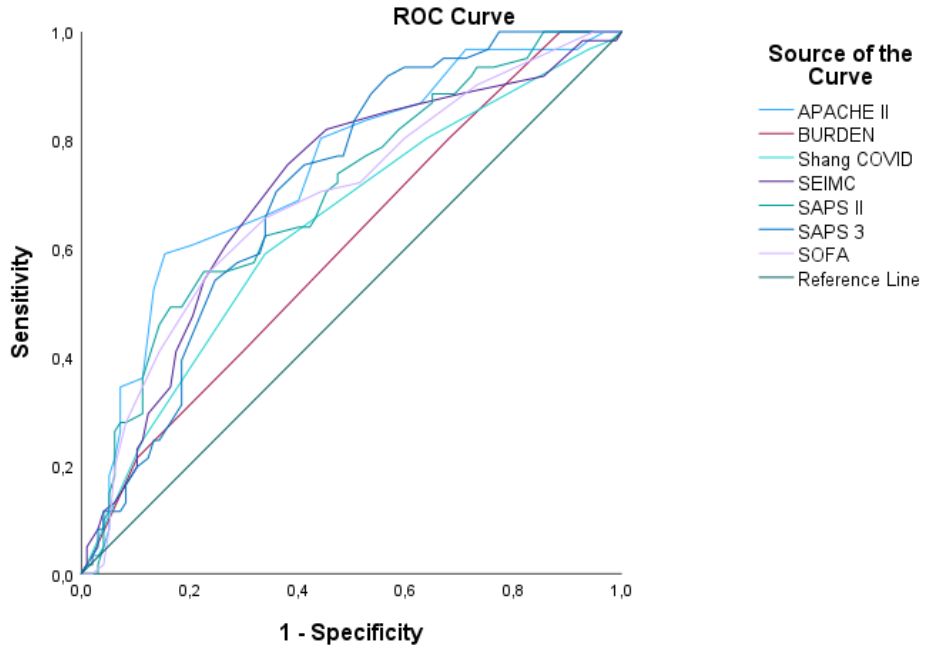

Figure 4.2.4.1 - Severity scores ROC curves for ICU mortality outcome (second wave).

Table 4.2.4.2 represents the estimated AUC result, p-value and CI for inflammation-based score in ICU mortality outcome (second wave).

Table 4.2.4.2 - Inflammation-based score estimated AUC, p-value and CI for ICU mortality outcome (second wave).

AUC estimate	p-value	95% Confidence Interval	
		Lower Bound	Upper Bound
0.485	0.785	0.381	0.590

Figure 4.2.4.2 represents inflammation-based score ROC curve (blue line) for ICU mortality outcome. As it can be seen, the line is very close to the reference line (red), meaning lack of discriminative ability.

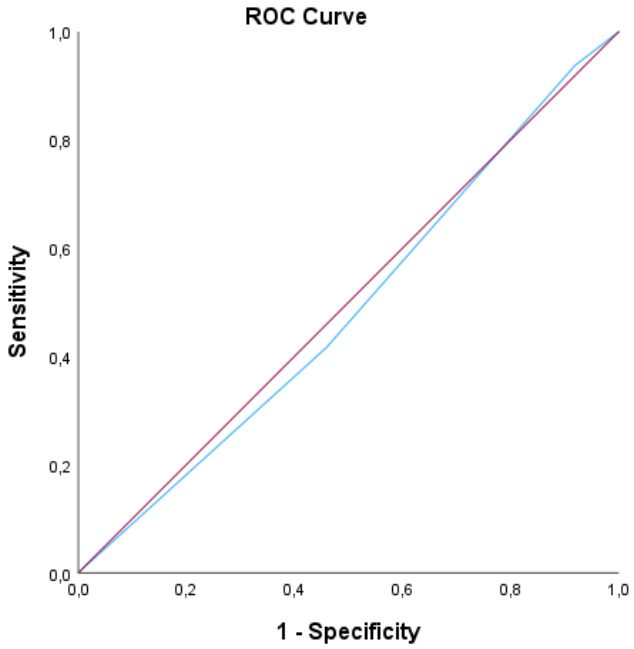

Figure 4.2.4.2 - Inflammation-based score ROC curve for ICU mortality outcome (second wave).

In Table 4.2.4.3 APACHE II demonstrates significantly higher discriminative ability compared to BURDEN (CI: 0.018; 0.262, p=0.025), indicating its superior performance in distinguishing between patient outcomes or conditions relative to this system. However, APACHE II shows no statistically significant differences when compared to other systems, although its comparison with SHANG approaches significance (CI: -0.002; 0.202, p=0.055).

BURDEN exhibits lower discriminative capacity in some comparisons. It shows significantly inferior performance compared to SEIMC (CI: -0.204; -0.004, p=0.042), and its comparison with SAPS 3 is borderline significant (CI: -0.221; <0.001, p=0.050). This pattern

suggests that BURDEN may be less effective in differentiating between patient outcomes in intensive care settings compared to some other scoring systems.

Notably, SHANG, SEIMC, SAPS II, SAPS 3, and SOFA do not show statistically significant differences in discriminative capacity when compared to each other. All confidence intervals for these comparisons include zero, and p-values are consistently above the conventional 0.05 threshold. This suggests that these scoring systems may have similar abilities to distinguish between different patient outcomes or conditions in intensive care units.

Table 4.2.4.3 - Results for paired-test for the difference in AUCs for ICU mortality outcome (second wave).

Scores	BURDEN	SHANG	SEIMC	SAPS II	SAPS 3	SOFA
APACHE II	(0.018; 0.262) p=0.025	(-0.002; 0.202) p=0.055	(-0.065; 0.136) p=0.485	(-0.031; 0.115) p=0.257	(-0.052; 0.110) p=0.481	(-0.041; 0.141) p=0.282
BURDEN		(-0.152; 0.072) p=0.485	(-0.204; -0.004) p=0.042	(-0.219; 0.024) p=0.117	(-0.221; <0.001) p=0.050	(-0.205; 0.026) p=0.127
SHANG			(-0.144; 0.016) p=0.116	(-0.156; 0.040) p=0.249	(-0.163; 0.022) p=0.135	(-0.161; 0.062) p=0.383
SEIMC				(-0.090; 0.103) p=0.895	(-0.095; 0.082) p=0.885	(-0.092; 0.121) p=0.792
SAPS II					(-0.094; 0.068) p=0.753	(-0.068; 0.084) p=0.840
SAPS 3						(-0.058; 0.100) p=0.606

Paired test for the difference in AUCs. Each result corresponds to 95% confidence interval for the AUC difference and corresponding P-value.

4.2.5. Early mortality at ICU

Table 4.2.5.1 displays the AUC (area under the curve), p-values, and 95% confidence intervals for various scoring systems (APACHE II, SAPS II, SAPS 3, SOFA, SEIMC, SHANG, and BURDEN) used to predict patient outcomes.

Among these, SAPS 3 achieves the highest AUC at 0.822, with a p-value of <0.001 and a confidence interval from 0.748 to 0.896, indicating strong predictive accuracy. SEIMC follows closely with an AUC of 0.800 ($p < 0.001$), and a confidence interval between 0.703 and 0.896, suggesting that it performs almost as well as SAPS 3.

APACHE II and SOFA also demonstrate good predictive capability, with AUCs of 0.775 and 0.762, respectively (both with p-values <0.001). Their confidence intervals (APACHE II: 0.666–0.883, SOFA: 0.653–0.870) show that they are strong predictors, though they fall slightly below SAPS 3 and SEIMC.

SHANG, with an AUC of 0.736 ($p < 0.001$) and a confidence interval from 0.627 to 0.846, shows fair predictive ability, although it does not perform as strongly as SAPS 3, SEIMC, APACHE II, or SOFA. SAPS II has an AUC of 0.716, a p-value of 0.001, and a confidence interval from 0.599 to 0.834, indicating moderate predictive accuracy.

Finally, BURDEN has the lowest AUC at 0.619, with a p-value of 0.072 and a confidence interval from 0.500 to 0.738, suggesting limited predictive performance, nearing random chance.

In summary, SAPS 3 and SEIMC emerge as the most accurate predictors, closely followed by APACHE II and SOFA. SHANG and SAPS II provide moderate predictive accuracy, while BURDEN shows the least reliability in predictive performance.

Table 4.2.5.1 - Severity scores estimated AUC, p-value and CI for early mortality at ICU outcome (second wave).

Scores	AUC estimate	p-value	95% Confidence Interval	
			Lower Bound	Upper Bound
APACHE II	0.775	<0.001	0.666	0.883
SAPS II	0.716	0.001	0.599	0.834
SAPS 3	0.822	<0.001	0.748	0.896
SOFA	0.762	<0.001	0.653	0.870
SEIMC	0.800	<0.001	0.703	0.896
SHANG	0.736	<0.001	0.627	0.846
BURDEN	0.619	0.072	0.500	0.738

As it can be seen in figure 4.2.5.1, SAPS 3 demonstrates the highest discriminatory ability with an AUC of 0.822, suggesting its ROC curve bows strongly towards the upper-left corner of the plot. This indicates a high true positive rate (sensitivity) even at low false positive rates (high specificity). The SEIMC and APACHE II scores also show excellent performance with AUCs of 0.800 and 0.775 respectively, implying their ROC curves are also well-curved towards the upper-left corner, though not as pronounced as SAPS 3.

SOFA and SHANG scores present good discriminatory power with AUCs of 0.762 and 0.736 respectively. Their ROC curves show a clear separation from the diagonal line of no discrimination, indicating a good balance between sensitivity and specificity.

The SAPS II score, with an AUC of 0.716, while still demonstrating fair discriminatory ability, it presents a ROC curve that is less bowed compared to the top-performing scores. Its curve would still be noticeably above the diagonal line, but with a more gradual ascent. The

BURDEN score, with the lowest AUC of 0.619, has a ROC curve that is closest to the diagonal line among all scores presented. This suggests a limited ability to discriminate between positive and negative outcomes, and its curve would show only a slight bow above the line of no discrimination.

It's noteworthy that all scores except BURDEN have p-values <0.001, indicating that their ROC curves are significantly different from the diagonal line of no discrimination. The BURDEN score, with a p-value of 0.072, is not statistically significant at the conventional 0.05 level, suggesting its curve is not significantly different from random chance.

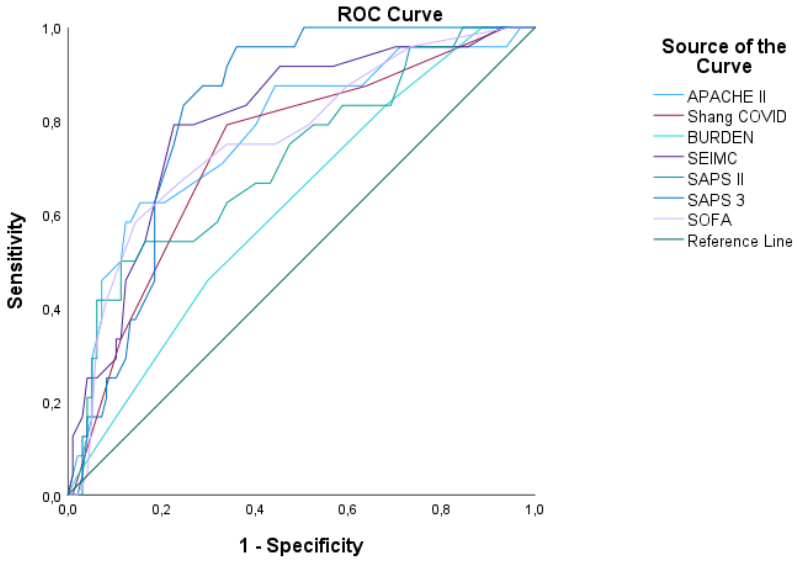

Figure 4.2.5.1 - Severity scores ROC curves for early ICU mortality, in second wave.

Table 4.2.5.2 represents the estimated AUC result, p-value and CI for inflammation-based score in early ICU mortality outcome (second wave).

Table 4.2.5.2 - Inflammation-based score estimated AUC, p-value and CI for early mortality at ICU outcome (second wave).

AUC estimate	p-value	95% Confidence Interval	
		Lower Bound	Upper Bound
0.413	0.267	0.265	0.561

Figure 4.2.5.2 represents inflammation-based score ROC curve (blue line) for early ICU mortality outcome. As it can be seen, the line is very close to the reference line (red), meaning poor discriminative ability.

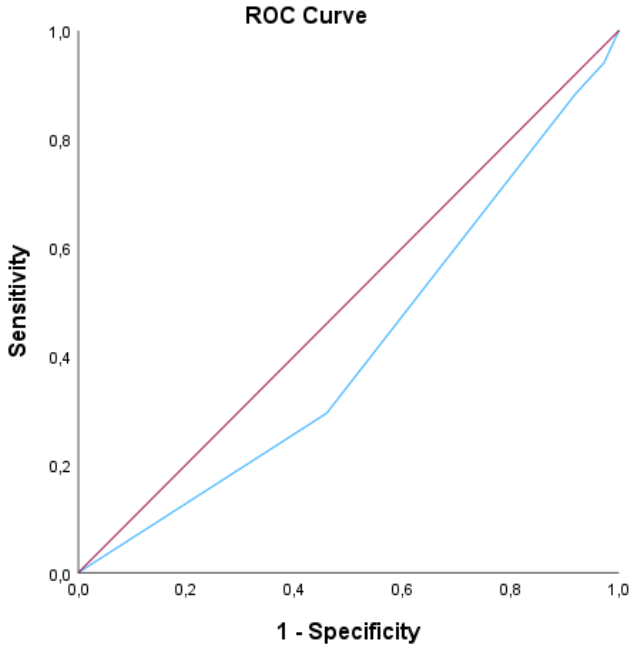

Figure 4.2.5.2 - Inflammation-based score ROC curve for early ICU mortality outcome (second wave).

As shown in Table 4.2.5.3, starting with APACHE II, the confidence intervals across its comparisons with other scores are notably wide, such as (-0.107; 0.185) when compared to SHANG, which indicates uncertainty about its relative effectiveness. The p-value of 0.604 further suggests that there is no statistically significant difference between APACHE II and SHANG, implying that APACHE II may not offer a strong advantage in discriminatory capacity. Similar trends are observed in comparisons with SOFA (p=0.858) and SAPS II (p=0.248), where the lack of significant differences raises questions about the utility of APACHE II in these contexts.

In contrast, SHANG shows mixed results. The CI for SHANG compared to BURDEN is (-0.026; 0.260) with a p-value of 0.108, suggesting a potential but non-significant difference in discriminatory capacity. However, when comparing SHANG to SEIMC, the CI of (-0.174; 0.047) indicates a more promising outcome, although the p-value remains non-significant at 0.260. This variability highlights the need for careful interpretation when considering SHANG's effectiveness.

The BURDEN score stands out with significant findings against SEIMC and SAPS 3. The CI for BURDEN compared to SEIMC is (-0.313; -0.048), which does not cross zero and has a p-value of 0.008, indicating that BURDEN is significantly less effective than SEIMC in predicting outcomes. Additionally, the comparison with SAPS 3 shows a significant negative difference (CI: -0.339; -0.066) and a p-value of 0.004, reinforcing concerns about BURDEN's reliability as a predictive tool.

When examining SEIMC, it shows non-significant differences against both SAPS II (p=0.223) and SOFA (p=0.646), suggesting that its discriminatory capacity is comparable but not superior to these scoring systems.

The SAPS II score demonstrates a significant finding against SAPS 3 with a CI of (-0.209; -0.002) and a p-value of 0.045, indicating that SAPS II may be more effective than SAPS 3 in certain contexts.

Finally, SAPS 3 shows non-significant results across its comparisons, including with SOFA (CI: -0.046; 0.166; p=0.266), suggesting limited discriminatory capacity relative to other scores.

Table 4.2.5.3 - Results for paired-test for the difference in AUCs for early ICU mortality outcome (second wave).

Scores	SHANG	BURDEN	SEIMC	SAPS II	SAPS 3	SOFA
APACHE II	(-0.107; 0.185) p=0.604	(-0.029; 0.340) p=0.098	(-0.169; 0.120) p=0.735	(-0.041; 0.157) p=0.248	(-0.151; 0.056) p=0.370	(-0.129; 0.154) p=0.858
SHANG		(-0.026; 0.260) p=0.108	(-0.174; 0.047) p=0.260	(-0.114; 0.153) p=0.774	(-0.191; 0.019) p=0.109	(-0.167; 0.115) p=0.720
BURDEN			(-0.313; -0.048) p=0.008	(-0.284; 0.089) p=0.306	(-0.339; -0.066) p=0.004	(-0.295; 0.010) p=0.067
SEIMC				(-0.051; 0.217) p=0.223	(-0.118; 0.073) p=0.646	(-0.113; 0.188) p=0.622
SAPS II					(-0.209; -0.002) p=0.045	(-0.160; 0.070) p=0.440
SAPS 3						(-0.046; 0.166) p=0.266

Paired test for the difference in AUCs. Each result corresponds to 95% confidence interval for the AUC difference and corresponding P-value.

4.2.6. Late mortality at ICU

As shown in Table 4.2.6.1, the APACHE II score achieves the highest AUC at 0.719, with a p-value less than 0.001 and a confidence interval of 0.624 to 0.814, indicating a high level of reliability and precision in prediction. In comparison, SAPS II has a slightly lower AUC of 0.686, with a p-value of 0.001 and a confidence interval of 0.587 to 0.784, suggesting it is a reasonably reliable predictor, though slightly less precise than APACHE II.

The SAPS 3 and SOFA scores have AUCs of 0.634 and 0.646, respectively, with p-values of 0.016 and 0.009. Their confidence intervals, ranging from 0.539 to 0.729 for SAPS 3 and 0.542 to 0.751 for SOFA, indicate moderate reliability in prediction, although not as strong as APACHE II or SAPS II. The SEIMC score also performs moderately, with an AUC of 0.641, a p-value of 0.011, and a confidence interval from 0.537 to 0.746, comparable to SAPS 3 and SOFA in accuracy.

In contrast, SHANG and BURDEN show weaker performance, with AUCs of 0.576 and 0.592, respectively. Their p-values of 0.173 and 0.101, along with broader confidence intervals (0.467 to 0.685 for SHANG and 0.486 to 0.697 for BURDEN), reflect lower reliability and precision in outcome prediction.

In summary, APACHE II demonstrates the highest predictive reliability among these scores, followed by SAPS II, which is also reasonably precise. SAPS 3, SOFA, and SEIMC show moderate predictive utility, while SHANG and BURDEN perform less effectively based on their AUCs and confidence intervals.

Table 4.2.6.1 - Severity scores estimated AUC, p-value and CI for late ICU mortality outcome (second wave).

Scores	AUC estimate	p-value	95% Confidence Interval	
			Lower Bound	Upper Bound
APACHE II	0.719	<0.001	0.624	0.814
SAPS II	0.686	0.001	0.587	0.784
SAPS 3	0.634	0.016	0.539	0.729
SOFA	0.646	0.009	0.542	0.751
SEIMC	0.641	0.011	0.537	0.746
SHANG	0.576	0.173	0.467	0.685

As shown in Figure 4.2.6.1, the ROC curve for APACHE II, with an AUC of 0.719, is well above the diagonal, demonstrating strong discriminatory power. The curve indicates that APACHE II is highly effective at distinguishing between the two categories, suggesting a high level of accuracy in its predictions.

SAPS II, with an AUC of 0.686, also displays a ROC curve above the diagonal, though its slope is slightly less steep than that of APACHE II. While still a reliable predictor, SAPS II shows a somewhat reduced ability to differentiate between outcomes compared to APACHE II, reflecting a slightly lower level of predictive accuracy.

The ROC curves for SAPS 3 and SOFA, with AUCs of 0.634 and 0.646, respectively, remain above the diagonal but are less pronounced than those of APACHE II and SAPS II. These curves indicate moderate discriminatory power, suggesting that while both scores are capable of distinguishing between outcomes, their predictive accuracy is lower than the top-performing systems.

SEIMC, with an AUC of 0.641, shows a ROC curve like those of SAPS 3 and SOFA. It reflects a moderate ability to distinguish between categories, but the curve is not as sharp as the higher-performing systems. This indicates a somewhat weaker predictive capability.

In contrast, SHANG and BURDEN, with AUCs of 0.576 and 0.592, respectively, show ROC curves closer to the diagonal. These curves indicate weaker discriminatory abilities, suggesting that both models are less effective at differentiating between outcomes. Their flatter slopes reflect their limited performance in terms of predictive accuracy.

In summary, APACHE II demonstrates the most robust ROC curve, with clear and reliable discrimination between outcomes. SAPS II is also strong, though slightly less effective. SAPS 3, SOFA, and SEIMC show moderate discriminatory power, while SHANG and BURDEN have weaker ROC curves, indicating less reliable predictions.

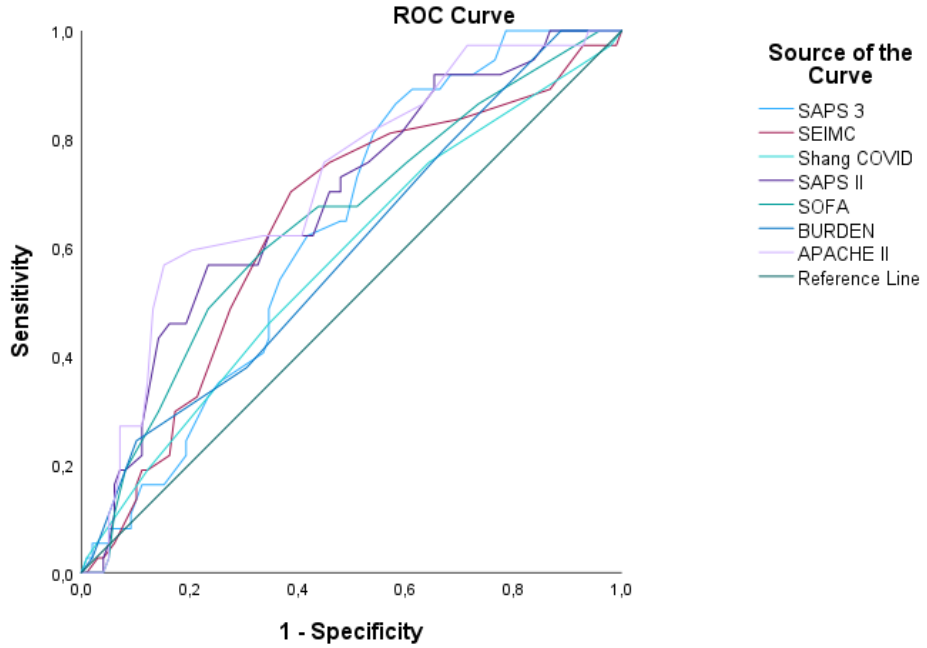

Figure 4.2.6.1 - Severity scores ROC curves for late ICU mortality outcome (second wave).

Table 4.2.6.2 represents the estimated AUC result, p-value and CI for inflammation-based score in late ICU mortality outcome (second wave).

Table 4.2.6.2 - Severity scores estimated AUCs, p-value and CI for late ICU mortality outcome (second wave).

AUC estimate	p-value	95% Confidence Interval	
		Lower Bound	Upper Bound
0.531	0.613	0.412	0.651

Figure 4.2.6.2 represents inflammation-based score ROC curve (blue line) for late ICU mortality outcome (second wave). As it can be seen, the line is very close to the reference line (red), meaning poor discriminative ability.

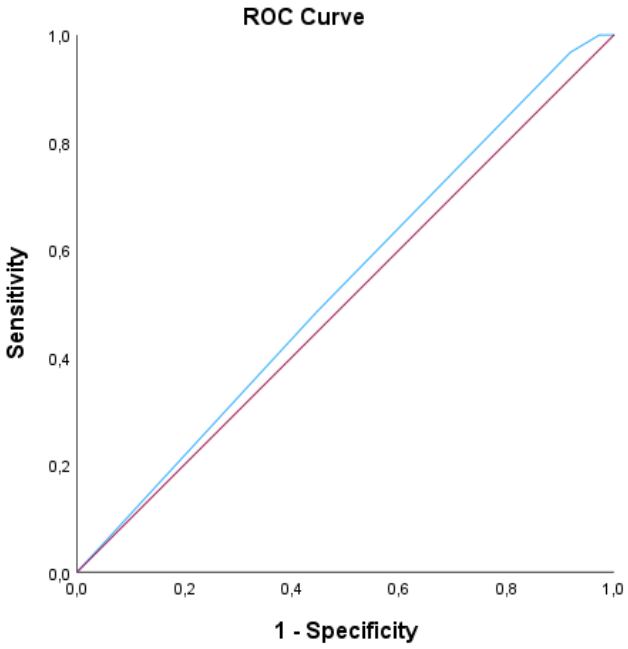

Figure 4.2.6.2 - Inflammation-based score ROC curve for late mortality outcome (second wave).

In Table 4.2.6.3, starting with SHANG, it shows varied results in its comparisons with other scoring systems. The confidence interval when compared to SEIMC is (-0.166; 0.036) with a p-value of 0.206, indicating no statistically significant difference in AUC between these two systems. However, the comparison with SAPS II yields a CI of (-0.225; 0.007) and a p-value of 0.064, suggesting that SHANG may be less effective than SAPS II, although this result approaches significance without reaching it. Comparisons with SAPS 3 (p=0.339), SOFA (p=0.336), and BURDEN (p=0.825) indicate no significant differences, reinforcing that SHANG does not demonstrate a clear advantage over these scoring systems.

SEIMC also exhibits mixed results across its comparisons. The CI against SAPS II is (-0.159; 0.070) with a p-value of 0.451, indicating no significant difference in AUC between these two scores. Similarly, comparisons with SAPS 3 (p=0.902) and SOFA (p=0.939) yield non-significant results, suggesting that SEIMC performs comparably to these systems in terms of discriminatory capacity. The comparison with APACHE II has a CI of (-0.198; 0.042) and a p-value of 0.202, which indicates that while there is no significant difference, the result is close to significance.

SAPS II demonstrates limited discriminatory capacity across its comparisons, with all p-values exceeding 0.05, indicating non-significant differences against SAPS 3 (p=0.310), SOFA (p=0.394), and APACHE II (p=0.069). The CI for SAPS II compared to SOFA is particularly wide (-0.046; 0.234), reflecting uncertainty about any potential differences in AUC.

When examining SAPS 3, it shows non-significant differences against SOFA (p=0.806) and APACHE II (p=0.080). The CI for SAPS 3 compared to APACHE II is (-0.181; 0.010), which approaches significance but ultimately does not reach it. The SOFA score exhibits non-significant results across its comparisons, particularly against APACHE II (p=0.187) and BURDEN (p=0.442). These findings suggest that SOFA does not provide a significant advantage over these other scoring systems in terms of AUC.

Finally, APACHE II has a CI of (-0.010; 0.266) when compared to BURDEN, with a p-value of 0.069, indicating that while there may be some difference in discriminatory capacity between APACHE II and BURDEN, it does not reach statistical significance at the conventional threshold of 0.05. The BURDEN score shows limited effectiveness overall, as indicated by its comparisons with other scoring systems yielding non-significant results across the board—most notably against SHANG (p=0.825), SEIMC (p=0.412), and others.

Table 4.2.6.3 - Results for paired-test for the difference in AUCs for late ICU mortality outcome (second wave).

Scores	SEIMC	SAPS II	SAPS 3	SOFA	APACHE II	BURDEN
SHANG	(-0.166; 0.036) p=0.206	(-0.225; 0.007) p=0.064	(-0.177; 0.061) p=0.339	(-0.213; 0.073) p=0.336	(-0.265; -0.022) p=0.021	(-0.151; 0.121) p=0.825
SEIMC		(-0.159; 0.070) p=0.451	(-0.109; 0.123) p=0.902	(-0.132; 0.122) p=0.939	(-0.198; 0.042) p=0.202	(-0.069; 0.169) p=0.412
SAPS II			(-0.048; 0.151) p=0.310	(-0.051; 0.129) p=0.394	(-0.129; 0.062) p=0.487	(-0.046; 0.234) p=0.189
SAPS 3				(-0.110; 0.086) p=0.806	(-0.181; 0.010) p=0.080	(-0.093; 0.178) p=0.537
SOFA					(-0.181; 0.035) p=0.187	(-0.085; 0.195) p=0.442
APACHE II						(-0.010; 0.266) p=0.069

Paired-test for the difference in AUCs. Each result corresponds to 95% confidence interval for the AUC difference and corresponding P-value.

4.2.7. Discussion of the results within second wave

As shown in Table 4.2.7.1 APACHE II shows strong performance across all categories. For Hospital Mortality, it has an AUC of 0.769, with a confidence interval ranging from 0.695 to 0.843, indicating reliable discriminatory ability. It also demonstrates solid performance in ICU Mortality (AUC = 0.742), Early ICU Mortality (AUC = 0.775), and Late ICU Mortality (AUC = 0.719), with all values showing good discriminatory power as indicated by their confidence intervals.

SAPS II performs well, though slightly lower than APACHE II in most categories. Its highest AUC is found in Hospital Mortality (0.755), followed by ICU Mortality (0.700), Early ICU Mortality (0.716), and Late ICU Mortality (0.686). While it remains a reliable predictor, the AUC values for SAPS II are somewhat lower, especially for Late ICU Mortality, where the discriminatory power is the weakest compared to the other outcomes.

SAPS 3 shows good discriminatory power for Hospital Mortality (AUC = 0.759) and ICU Mortality (AUC = 0.713), but its performance significantly drops for Late ICU Mortality (AUC =

0.634). However, it is particularly strong in predicting Early ICU Mortality, with an impressive AUC of 0.822. This suggests SAPS 3 is better suited for early outcome predictions.

SOFA has moderate performance across all outcomes. The highest AUC is observed for Early ICU Mortality (AUC = 0.762), followed by Hospital Mortality (AUC = 0.717). Its weakest performance is for Late ICU Mortality (AUC = 0.646), indicating less effective discrimination in predicting later-stage mortality. The SOFA score is used to assess organ dysfunction, which makes it more relevant for early mortality since organ failure typically occurs in the early stages of severe illness, before a fatal outcome. This score is often used to monitor critically ill patients, such as those in intensive care units. When used in situations with an immediate risk of death, the SOFA score can quickly identify clinical deterioration and provide guidance for urgent treatment. For this reason, SOFA is particularly effective in predicting early mortality, being an important indicator in the early phases of the disease when prompt intervention may be critical to saving lives.

SEIMC demonstrates strong performance in predicting Hospital Mortality (AUC = 0.724) and ICU Mortality (AUC = 0.706). However, it performs exceptionally well in Early ICU Mortality (AUC = 0.800), suggesting it is a more reliable predictor in the early stages. Its performance for Late ICU Mortality (AUC = 0.641) is relatively weaker.

SHANG has moderate to weak predictive power overall. It performs the best for Early ICU Mortality (AUC = 0.736), but its AUC for Hospital Mortality (0.648) and ICU Mortality (0.642) are relatively low, suggesting less accuracy in predicting these outcomes. Late ICU Mortality shows the weakest performance with an AUC of 0.576.

BURDEN consistently shows weak discriminatory ability. It has the lowest AUC for Hospital Mortality (0.564), followed by Late ICU Mortality (0.592), Early ICU Mortality (0.619), and ICU Mortality (0.602). The confidence intervals are wide, reflecting less precision and weaker reliability in predicting any of the mortality categories.

In summary, APACHE II and SAPS 3 demonstrate strong and reliable discriminatory power across all mortality categories, particularly for hospital and ICU mortality. SEIMC excels in early ICU mortality prediction, while SAPS II and SOFA show moderate discriminatory abilities. SHANG and BURDEN have limited performance, particularly in predicting Late ICU Mortality. These results suggest that the choice of scoring system may depend on the specific mortality category being predicted, with APACHE II and SAPS 3 generally offering the best overall predictive ability.

Table 4.2.7.1 - Estimated AUCs and confidence intervals results for mortality outcomes, in second wave.

Scores	Hospital Mortality	ICU Mortality	Early ICU Mortality	Late ICU Mortality
APACHE II	0.769 (0.695 - 0.843)	0.742 (0.663 - 0.821)	0.775 (0.666 - 0.883)	0.719 (0.624 - 0.814)
SAPS II	0.755 (0.680 - 0.830)	0.700 (0.617 - 0.783)	0.716 (0.599 - 0.834)	0.686 (0.587 - 0.784)
SAPS 3	0.759 (0.683 - 0.834)	0.713 (0.634 - 0.792)	0.822 (0.748 - 0.896)	0.634 (0.539 - 0.729)
SOFA	0.717 (0.638 - 0.797)	0.692 (0.607 - 0.777)	0.762 (0.653 - 0.870)	0.646 (0.542 - 0.751)
SEIMC	0.724 (0.642 - 0.805)	0.706 (0.622 - 0.790)	0.800 (0.703 - 0.896)	0.641 (0.537 - 0.746)
SHANG	0.648 (0.561 - 0.734)	0.642 (0.553 - 0.731)	0.736 (0.627 - 0.846)	0.576 (0.467 - 0.685)
BURDEN	0.564 (0.474 - 0.654)	0.602 (0.513 - 0.691)	0.619 (0.500 – 0.738)	0.592 (0.486 - 0.697)

As more data on the virus became available through global and local studies, treatment strategies shifted. By the second wave, therapies became more evidence based. For example, the use of corticosteroids, such as dexamethasone, was better understood, and it became the standard for severe cases due to its ability to reduce inflammation and improve outcomes in critically ill patients. Other antiviral agents, such as remdesivir, continued to be used, though with more clarity on their specific indications, (INFARMED - Autoridade Nacional do Medicamento e Produtos de Saúde, I.P., 2022).

The application of inflammatory markers for disease severity assessment, like the SOFA score, initially served as a key tool for predicting outcomes in COVID-19 patients. However, in later waves, with the more widespread use of anti-inflammatory drugs (such as dexamethasone), these markers became less reliable. As these treatments reduced the inflammatory response, they could mask the true severity of disease, leading to less accurate predictions based on such scores. This reduction in predictive power over time suggests that changes in therapeutic strategies, particularly anti-inflammatory treatments, likely influenced the utility of these inflammatory-based scores, (INFARMED - Autoridade Nacional do Medicamento e Produtos de Saúde, I.P., 2022).

4.3. Scores comparison between first and second waves of COVID-19

In this section, we compare the performance of various clinical scoring systems between the first and second waves of COVID-19. This comparison aims to assess how well

these scores predicted patient outcomes across different phases of the pandemic, potentially highlighting shifts in patient severity, healthcare responses, or the effectiveness of scoring systems in adapting to the evolving clinical characteristics of COVID-19 over time. When discussing COVID-19-specific scores, it is important to note that these scores were validated with populations exposed to the virus during the first wave of the pandemic.

The first wave was marked by the emergence of the SARS-CoV-2 virus, and the populations involved had different demographic, clinical, and genetic characteristics. The loss of efficacy of these scores in the second wave can partly be explained by changes in the population and disease characteristics. Additionally, the evolution of the virus and the introduction of new variants could have altered the disease's progression, making triage tools less accurate.

4.3.1. Scores comparison between first and second waves for hospital mortality outcome

The results presented in Table 4.3.1.1 compare the performance of various scoring systems (APACHE II, SAPS II, SAPS 3, SOFA, SEIMC, SHANG, BURDEN, and INFLAMMATION-BASED) by showing the AUC (Area Under the Curve) for both Wave 1 and Wave 2, along with the p-value that assesses the statistical significance of the difference in performance between the two waves.

For APACHE II, the AUC in Wave 1 is 0.704 (95% CI: 0.606–0.803), and it increases to 0.769 (95% CI: 0.695–0.843) in Wave 2. The p-value of 0.301 suggests that there is no statistically significant difference between the two waves, meaning the improvement in performance from Wave 1 to Wave 2 is not large enough to be considered meaningful.

SAPS II shows an AUC of 0.685 (95% CI: 0.579–0.790) in Wave 1, which increases slightly to 0.755 (95% CI: 0.680–0.830) in Wave 2. The p-value of 0.288 also indicates that the difference between the two waves is not statistically significant, suggesting that while there is an improvement, it is not substantial enough to be considered a meaningful change.

For SAPS 3, the AUC in Wave 1 is 0.735 (95% CI: 0.634–0.835), and in Wave 2, it increases to 0.763 (95% CI: 0.689–0.837). The p-value of 0.656 shows no significant difference between the two waves, indicating that SAPS 3 maintains a similar performance across both waves.

In the case of SOFA, the AUC in Wave 1 is 0.680 (95% CI: 0.571–0.788), and it improves to 0.718 (95% CI: 0.639–0.797) in Wave 2. However, the p-value of 0.574 shows that the difference is not statistically significant, suggesting a modest but not significant improvement in predictive power between the two waves.

SEIMC demonstrates the highest AUC in Wave 1 (0.810, 95% CI: 0.725–0.895), but this drops to 0.723 (95% CI: 0.643–0.803) in Wave 2. The p-value of 0.146 indicates that the

change in AUC is not statistically significant, meaning the decrease in performance from Wave 1 to Wave 2 is not large enough to be considered meaningful.

For SHANG, the AUC in Wave 1 is 0.724 (95% CI: 0.624–0.824), and it decreases to 0.650 (95% CI: 0.563–0.736) in Wave 2. The p-value of 0.272 suggests that the difference in performance between the two waves is not statistically significant, reflecting a moderate drop in predictive power between the waves.

BURDEN shows the weakest performance with an AUC of 0.505 (95% CI: 0.384–0.625) in Wave 1, and a slight increase to 0.567 (95% CI: 0.478–0.656) in Wave 2. The p-value of 0.413 indicates no significant difference between the two waves, meaning that BURDEN remains a weak predictor in both waves.

Finally, INFLAMMATION-BASED shows an AUC of 0.583 (95% CI: 0.452–0.713) in Wave 1, and a slight drop to 0.483 (95% CI: 0.380–0.586) in Wave 2. The p-value of 0.241 indicates no significant difference between the two waves, suggesting that this scoring system is consistently weak across both waves.

In conclusion, while some scores, such as APACHE II, SAPS II, and SAPS 3, show slight improvements in AUC from Wave 1 to Wave 2, none of these changes are statistically significant, as reflected by the p-values greater than 0.05. SEIMC, while performing well in Wave 1, shows a decrease in Wave 2, but again, the difference is not statistically significant. SHANG and BURDEN consistently demonstrate weaker performance, and no meaningful improvement is observed between the waves. Therefore, while some scores show minor changes in AUC, these differences are not significant, and the overall predictive ability of these scoring systems remains relatively stable across the two waves.

Table 4.3.1.1 - Comparison of severity scores between first and second waves for hospital mortality.

Scores	Wave	AUC estimate	95% Confidence Interval		p-value
			Lower Bound	Upper Bound	
APACHE II	1	0.704	0.606	0.803	0.301
	2	0.769	0.695	0.843	
SAPS II	1	0.685	0.579	0.790	0.288
	2	0.755	0.680	0.830	
SAPS 3	1	0.735	0.634	0.835	0.656
	2	0.763	0.689	0.837	
SOFA	1	0.680	0.571	0.788	0.574

	2	0.718	0.639	0.797	
SEIMC	1	0.810	0.725	0.895	0.146
	2	0.723	0.643	0.803	
SHANG	1	0.724	0.624	0.824	0.272
	2	0.650	0.563	0.736	
BURDEN	1	0.505	0.384	0.625	0.413
	2	0.567	0.478	0.656	
INFLAMMATION-BASED	1	0.583	0.452	0.713	0.241
	2	0.483	0.380	0.586	

Figure 4.3.1.1 represents the ROC curve of each score across both waves. For APACHE II, the ROC curve in Wave 1 shows moderate performance with an AUC of 0.704, which improves to 0.769 in Wave 2. However, despite this increase, the ROC curve for APACHE II still does not show a statistically significant difference between the waves, as indicated by the p-value of 0.301. The curve in both waves is above the diagonal, but the improvement in performance is not marked enough to be considered meaningful.

SAPS II also shows an increase in its ROC curve performance from Wave 1 (AUC = 0.685) to Wave 2 (AUC = 0.755), but the change is not statistically significant (p-value = 0.288). The ROC curve in both waves reflects a reliable ability to distinguish outcomes, though the improvement is subtle and not substantial enough to justify a significant shift.

SAPS 3 demonstrates slightly better performance than SAPS II in both waves, with an AUC of 0.735 in Wave 1 and 0.763 in Wave 2. The ROC curve in both waves reflects good discriminatory power, but again, the difference is not statistically significant, as indicated by the p-value of 0.656.

SOFA shows a modest improvement in its ROC curve, with an AUC increasing from 0.680 in Wave 1 to 0.718 in Wave 2. While the ROC curve in both waves remains above the diagonal, the p-value of 0.574 suggests that the increase in performance is not statistically significant, implying only a slight, non-meaningful enhancement in predictive ability.

SEIMC has the highest AUC in Wave 1 (0.810), with a decrease to 0.723 in Wave 2. Despite the drop, the ROC curve in both waves indicates a strong ability to distinguish between outcomes, although the change is not statistically significant (p-value = 0.146). The curve suggests that SEIMC performs relatively well but shows a slight reduction in discriminative ability from Wave 1 to Wave 2.

For SHANG, the ROC curve starts at 0.724 in Wave 1 and drops to 0.650 in Wave 2. This decline reflects a reduced ability to discriminate between outcomes, with the p-value of 0.272 showing that the decrease is not statistically significant. The ROC curve in both waves

indicates that SHANG struggles to effectively distinguish between positive and negative outcomes.

BURDEN has the weakest ROC curve performance, with an AUC of 0.505 in Wave 1 and a slight increase to 0.567 in Wave 2. The ROC curves for BURDEN in both waves are close to the diagonal, suggesting a poor ability to discriminate between the outcomes. The p-value of 0.413 indicates that there is no meaningful difference between the waves.

INFLAMMATION-BASED shows an AUC of 0.583 in Wave 1 and a slight decline to 0.483 in Wave 2. The ROC curve for this scoring system is closer to the diagonal, indicating limited discriminatory power, and the p-value of 0.241 suggests no statistically significant change between the two waves.

In summary, while some scoring systems show slight improvements in their ROC curves from Wave 1 to Wave 2, such as APACHE II, SAPS II, and SAPS 3, these improvements are not substantial enough to be considered statistically significant. SEIMC, although initially performing well in Wave 1, shows a decrease in Wave 2, but this change is also not significant. SHANG and BURDEN consistently demonstrate weak discriminatory power, as reflected in their ROC curves, with little improvement observed across the waves. The overall discriminatory ability of these scores remains relatively stable, with few significant changes in performance between the two waves.

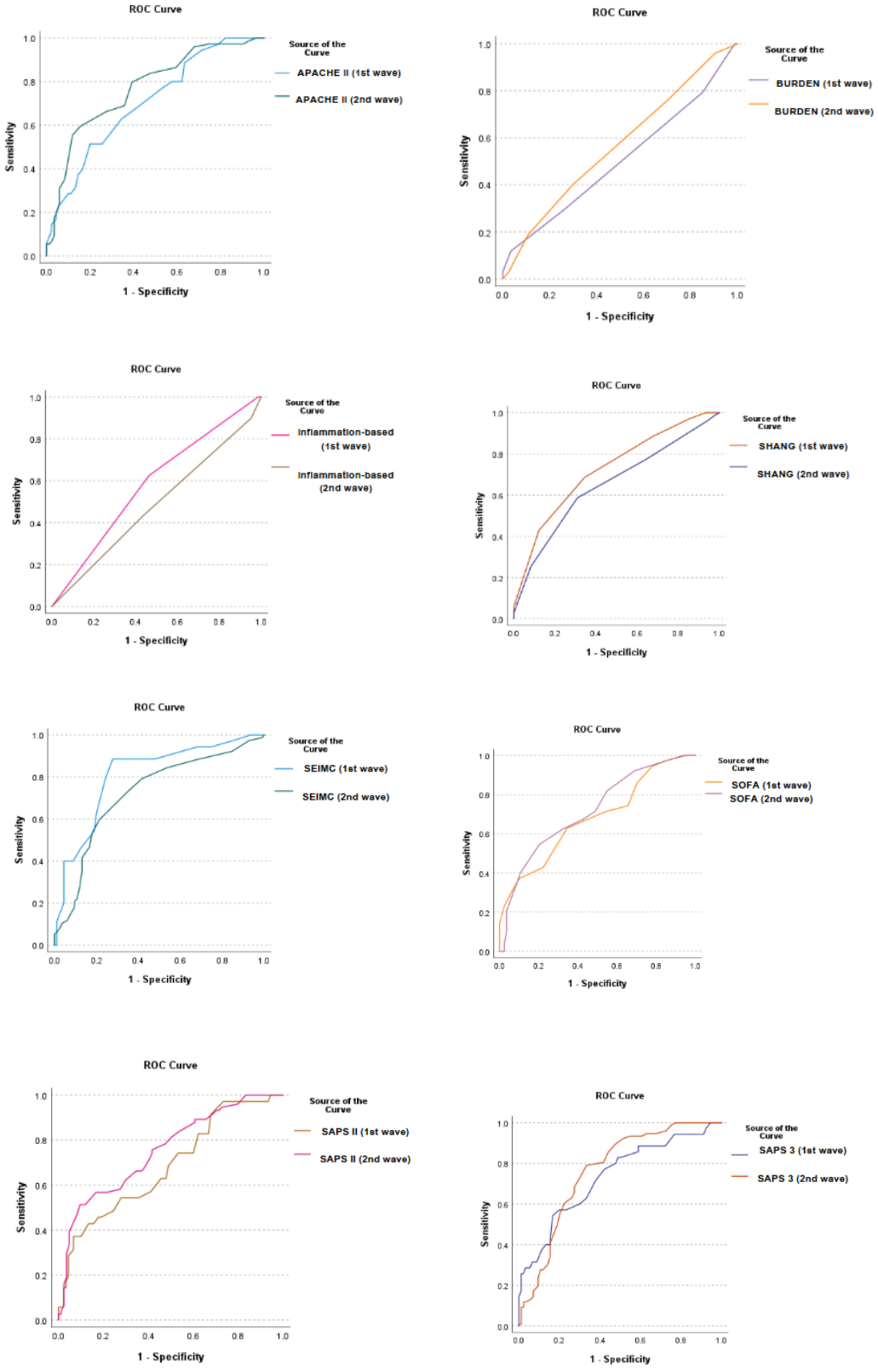

Figure 4.3.1.1 – AUC comparison for each score in first and second waves for hospital mortality outcome.

4.3.2. Scores comparison between first and second waves for ICU mortality

For each score, the Table 4.3.2.1 shows the performance of various scoring systems (APACHE II, SAPS II, SAPS 3, SOFA, SEIMC, SHANG, BURDEN, and INFLAMMATION-BASED) across two waves, with AUC values, 95% confidence intervals, and p-values.

APACHE II performs consistently well across both waves, with an AUC of 0.744 in Wave 1 and 0.742 in Wave 2. These values suggest strong predictive accuracy, with narrow confidence intervals indicating precision. The p-value of 0.973 indicates no statistically significant difference in performance between the two waves.

SAPS II shows slightly lower AUC values, with 0.718 in Wave 1 and 0.700 in Wave 2, still reflecting reasonable predictive ability. The confidence intervals overlap, indicating similar performance between the waves. The p-value of 0.800 suggests no significant change in the model's performance across the two waves.

SAPS 3 shows a stronger AUC of 0.781 in Wave 1 but declines to 0.719 in Wave 2. The confidence intervals for both waves indicate moderate to strong discriminatory power, although there is a decrease in the model's predictive ability in Wave 2. The p-value of 0.332 indicates that the change in performance between the waves is not statistically significant.

SOFA shows a decrease in AUC from 0.734 in Wave 1 to 0.692 in Wave 2. Despite the decline, both AUCs still reflect good discriminatory power. The p-value of 0.551 suggests that the difference in performance between the waves is not statistically significant.

SEIMC performs the best in Wave 1 with an AUC of 0.808, but this drops to 0.705 in Wave 2. Although the predictive ability decreases, the model still shows moderate performance in both waves. The p-value of 0.087 indicates a trend towards a decrease in performance, but it is not statistically significant.

SHANG shows weaker performance, with AUCs of 0.698 in Wave 1 and 0.644 in Wave 2. These values are still above the diagonal, but the model's predictive ability is weaker than others. The p-value of 0.456 suggests no significant difference between the two waves, indicating stable but limited predictive power.

BURDEN has the lowest AUC values, with 0.579 in Wave 1 and 0.605 in Wave 2, both close to the diagonal. These values indicate weak discriminatory power, and the p-value of 0.748 shows no significant difference between the two waves, confirming the model's limited reliability.

INFLAMMATION-BASED shows the weakest performance overall, with AUC values of 0.549 in Wave 1 and 0.485 in Wave 2. Both values are close to the diagonal, indicating poor discriminatory power. The p-value of 0.484 suggests no significant difference in performance between the two waves, highlighting the model's poor predictive capability.

In summary, APACHE II and SAPS 3 exhibit strong performance with consistent results across both waves, while SAPS II, SOFA, and SEIMC show moderate predictive ability with slight declines in Wave 2 but no significant changes. SHANG, BURDEN, and INFLAMMATION-BASED perform poorly, with no significant differences between the waves, reinforcing their limited ability to predict outcomes.

Table 4.3.2.1 - Comparison of severity scores between first and second waves for ICU mortality.

Scores	Wave	AUC estimate	95% Confidence Interval		p-value
			Lower Bound	Upper Bound	
APACHE II	1	0.744	0.641	0.847	0.973
	2	0.742	0.663	0.821	
SAPS II	1	0.718	0.604	0.832	0.800
	2	0.700	0.617	0.783	
SAPS 3	1	0.781	0.682	0.880	0.332
	2	0.719	0.641	0.797	
SOFA	1	0.734	0.626	0.842	0.551
	2	0.692	0.608	0.776	
SEIMC	1	0.808	0.724	0.893	0.087
	2	0.705	0.622	0.788	
SHANG	1	0.698	0.588	0.808	0.456
	2	0.644	0.556	0.732	
BURDEN	1	0.579	0.446	0.712	0.748
	2	0.605	0.517	0.694	
INFLAMMATION-BASED	1	0.549	0.405	0.692	0.484
	2	0.485	0.381	0.590	

APACHE II maintains a strong ROC curve across both waves, with AUC values of 0.744 in Wave 1 and 0.742 in Wave 2. These values indicate that the curve is well above the diagonal, reflecting good discriminatory power. The ROC curve shows a high ability to differentiate between the two categories, with minimal fluctuation between the two waves.

SAPS II also exhibits a reasonably strong ROC curve, with AUC values of 0.718 in Wave 1 and 0.700 in Wave 2. While the curve remains above the diagonal, there is a slight decrease in its steepness between waves, suggesting a slight reduction in predictive power, but still maintaining moderate accuracy.

SAPS 3 shows a higher AUC in Wave 1 (0.781) compared to Wave 2 (0.719), indicating that the ROC curve is initially stronger but becomes slightly flatter in the second wave. The overall shape of the curve in both waves still indicates reasonable discriminatory ability, although the decline suggests reduced effectiveness in the second wave.

SOFA's ROC curve is similar, with a decrease in AUC from 0.734 in Wave 1 to 0.692 in Wave 2. This decrease suggests that the curve in Wave 1 is more pronounced, providing better differentiation between the two categories. The curve in Wave 2 remains above the diagonal but is less steep, reflecting a decrease in discriminatory power.

SEIMC shows the strongest ROC curve in Wave 1 with an AUC of 0.808, indicating a curve well above the diagonal and strong predictive ability. However, the ROC curve flattens somewhat in Wave 2, with an AUC of 0.705. Despite the drop, it still reflects moderate discriminatory power, though the performance is not as strong as in Wave 1.

SHANG's ROC curve is weaker, with AUC values of 0.698 in Wave 1 and 0.644 in Wave 2. Both curves are above the diagonal but show moderate, less pronounced slopes. This indicates limited discriminatory power, with the curve in Wave 2 being even less effective at differentiating between the two categories.

BURDEN's ROC curve shows the weakest performance, with AUC values of 0.579 in Wave 1 and 0.605 in Wave 2. The curve is closer to the diagonal in both waves, indicating poor discriminatory ability and a tendency toward random classification.

INFLAMMATION-BASED has the weakest ROC curve overall, with AUC values of 0.549 in Wave 1 and 0.485 in Wave 2. Both curves are very close to the diagonal, suggesting that the model struggles to distinguish between categories in both waves, with the curve in Wave 2 reflecting even weaker performance.

In summary, APACHE II stands out with consistently strong ROC curves, while SAPS 3, SAPS II, and SOFA show moderate performance with slight declines in Wave 2. SEIMC shows a strong curve in Wave 1 but a decrease in Wave 2. SHANG, BURDEN, and INFLAMMATION-BASED exhibit weaker ROC curves, with limited discriminatory power in both waves.

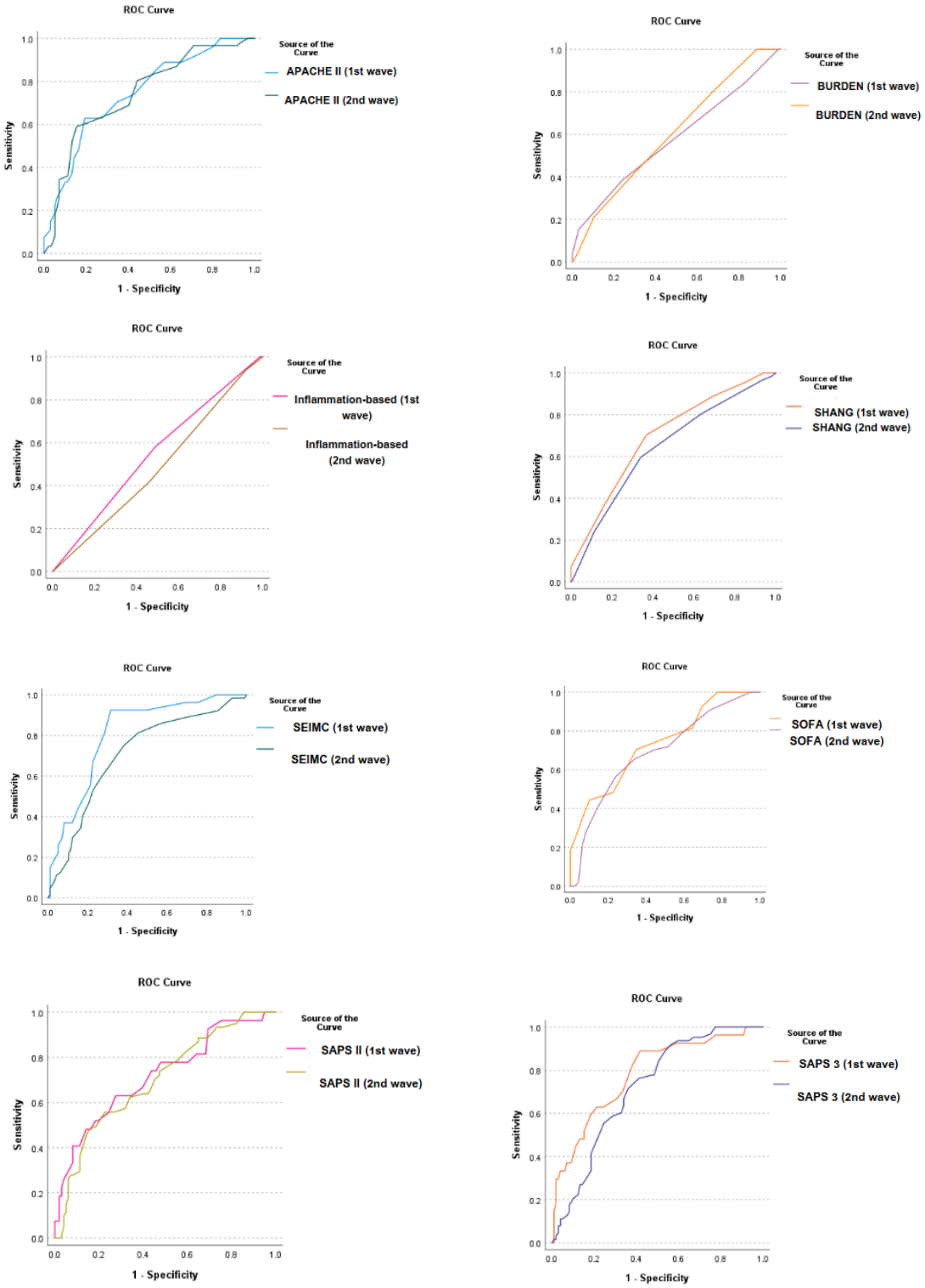

Figure 4.3.2.1 - AUC comparison for each score in first and second waves for ICU mortality outcome.

4.3.3. Scores comparison between first and second waves for early mortality at ICU

As presented in Table 4.3.3.1 APACHE II demonstrates strong performance in both waves, with AUC values of 0.747 in Wave 1 and 0.775 in Wave 2. The confidence intervals (Wave 1: 0.602–0.891, Wave 2: 0.666–0.883) indicate that the model consistently exhibits a reliable ability to differentiate between outcomes. The p-value of 0.762 suggests no significant difference between the two waves, implying that the performance of the APACHE II score remains stable across the two time points.

SAPS II also shows reasonable performance, with AUC values of 0.737 in Wave 1 and 0.716 in Wave 2. The confidence intervals (Wave 1: 0.572–0.902, Wave 2: 0.599–0.834) indicate reliable predictive accuracy, although the slightly lower AUC in Wave 2 suggests a mild decrease in discriminatory power. The p-value of 0.843 indicates that there is no significant difference in performance between the two waves.

SAPS 3 demonstrates a noticeable improvement between the two waves, with an AUC of 0.733 in Wave 1 and 0.828 in Wave 2. The confidence intervals (Wave 1: 0.592–0.875, Wave 2: 0.757–0.900) highlight a significant increase in discriminatory power in the second wave. However, the p-value of 0.240 indicates no significant difference between the two waves, suggesting that the observed improvement may not be statistically significant.

SOFA shows consistent performance across both waves, with AUC values of 0.756 in Wave 1 and 0.755 in Wave 2. The confidence intervals (Wave 1: 0.617–0.894, Wave 2: 0.650–0.860) confirm the model's stable predictive accuracy, and the p-value of 0.991 indicates no significant difference between the two waves, suggesting stable performance.

SEIMC displays high predictive power with AUC values of 0.783 in Wave 1 and 0.787 in Wave 2, accompanied by confidence intervals (Wave 1: 0.667–0.899, Wave 2: 0.694–0.880) that indicate strong discriminatory ability in both waves. The p-value of 0.964 suggests no significant difference between the waves, reflecting consistent predictive accuracy.

SHANG shows a weaker performance, with AUC values of 0.611 in Wave 1 and 0.738 in Wave 2. The confidence intervals (Wave 1: 0.445–0.777, Wave 2: 0.631–0.844) suggest that the model performs moderately well in Wave 2, though the AUC in Wave 1 is relatively low. The p-value of 0.210 suggests no significant difference between the two waves.

BURDEN exhibits relatively low predictive power, with AUC values of 0.573 in Wave 1 and 0.626 in Wave 2. The confidence intervals (Wave 1: 0.389–0.758, Wave 2: 0.510–0.743) indicate limited discriminatory ability, with both waves showing weak performance. The p-value of 0.635 indicates no significant difference in performance between the two waves.

INFLAMMATION-BASED shows the weakest performance overall, with AUC values of 0.537 in Wave 1 and 0.413 in Wave 2. The confidence intervals (Wave 1: 0.340–0.735, Wave

2: 0.265–0.561) reflect low discriminatory power, and the p-value of 0.325 suggests no significant difference in performance between the two waves.

In summary, APACHE II, SAPS II, SOFA, and SEIMC show stable and relatively strong performance across both waves. SAPS 3 improves slightly in Wave 2 but does not show a statistically significant difference. SHANG, BURDEN, and INFLAMMATION-BASED exhibit weaker performance, with INFLAMMATION-BASED showing the least predictive power across both waves. The p-values across the scores indicate that there are no significant differences between the two waves for most models, suggesting that their predictive abilities remain stable over time.

Table 4.3.3.1 - Comparison of severity scores between first and second waves for early mortality outcome.

Scores	Wave	AUC estimate	95% Confidence Interval		p-value
			Lower Bound	Upper Bound	
APACHE II	1	0.747	0.602	0.891	0.762
	2	0.775	0.666	0.883	
SAPS II	1	0.737	0.572	0.902	0.843
	2	0.716	0.599	0.834	
SAPS 3	1	0.733	0.592	0.875	0.240
	2	0.828	0.757	0.900	
SOFA	1	0.756	0.617	0.894	0.991
	2	0.755	0.650	0.860	
SEIMC	1	0.783	0.667	0.899	0.964
	2	0.787	0.694	0.880	
SHANG	1	0.611	0.445	0.777	0.210
	2	0.738	0.631	0.844	
BURDEN	1	0.573	0.389	0.758	0.635
	2	0.626	0.510	0.743	
INFLAMMATION-BASED	1	0.537	0.340	0.735	0.325
	2	0.413	0.265	0.561	

As shown in figure 4.3.3.1, APACHE II consistently performs well in both waves, with AUC values of 0.747 in Wave 1 and 0.775 in Wave 2. Both curves lie well above the diagonal, reflecting strong discriminatory power and a reliable ability to distinguish between outcomes.

The AUC values are quite similar between the two waves, indicating stable performance over time.

SAPS II also shows solid performance, with an AUC of 0.737 in Wave 1 and 0.716 in Wave 2. The ROC curve for SAPS II remains above the diagonal in both waves, suggesting it maintains moderate discriminatory ability, though slightly less effective in Wave 2 compared to Wave 1. The curve in Wave 2 is slightly flatter, suggesting a minor decrease in performance.

SAPS 3 displays a stronger ROC curve in Wave 2, with an AUC of 0.828 compared to 0.733 in Wave 1. This increase in AUC reflects improved discriminatory power, with the curve in Wave 2 being steeper and further from the diagonal, suggesting better performance in distinguishing between the outcomes. The curve in Wave 1 is more moderate, indicating moderate discriminatory ability, but the shift to a stronger curve in Wave 2 indicates potential improvements.

SOFA has relatively stable ROC curves between the two waves, with AUCs of 0.756 in Wave 1 and 0.755 in Wave 2. The curves are consistently above the diagonal, indicating stable discriminatory power. There is little to no change in the steepness of the curves between the waves, suggesting that SOFA's performance remains consistent over time.

SEIMC shows high discriminatory ability with AUCs of 0.783 in Wave 1 and 0.787 in Wave 2. Both ROC curves are well above the diagonal, and the curves are relatively steep in both waves, reflecting strong and consistent discriminatory power.

SHANG exhibits weaker ROC curves, with AUCs of 0.611 in Wave 1 and 0.738 in Wave 2. The ROC curve in Wave 1 is closer to the diagonal, suggesting weaker discriminatory ability. However, in Wave 2, the curve is more distinct from the diagonal, indicating improved performance in distinguishing between outcomes.

BURDEN shows even weaker ROC curves, with AUCs of 0.573 in Wave 1 and 0.626 in Wave 2. Both curves are near the diagonal, reflecting poor discriminatory power. The curve in Wave 2 is slightly better, but both waves show limited ability to distinguish between positive and negative outcomes.

INFLAMMATION-BASED has the weakest ROC curves, with AUCs of 0.537 in Wave 1 and 0.413 in Wave 2. The ROC curve in Wave 1 is slightly above the diagonal, but it is relatively flat, indicating poor ability to discriminate between outcomes. The curve in Wave 2 is even closer to the diagonal, reflecting poor discriminatory performance.

In summary, the APACHE II, SAPS II, SOFA, and SEIMC scores all display ROC curves that are well above the diagonal in both waves, indicating strong and consistent discriminatory power. SAPS 3 improves its discriminatory power in Wave 2, with a stronger ROC curve. SHANG, BURDEN, and INFLAMMATION-BASED show weaker ROC curves, with performance improving slightly for SHANG and BURDEN in Wave 2, though they remain relatively ineffective compared to the other models. Overall, the ROC curves reflect the

predictive accuracy of each scoring system, with SEIMC and APACHE II exhibiting the most robust performance.

.

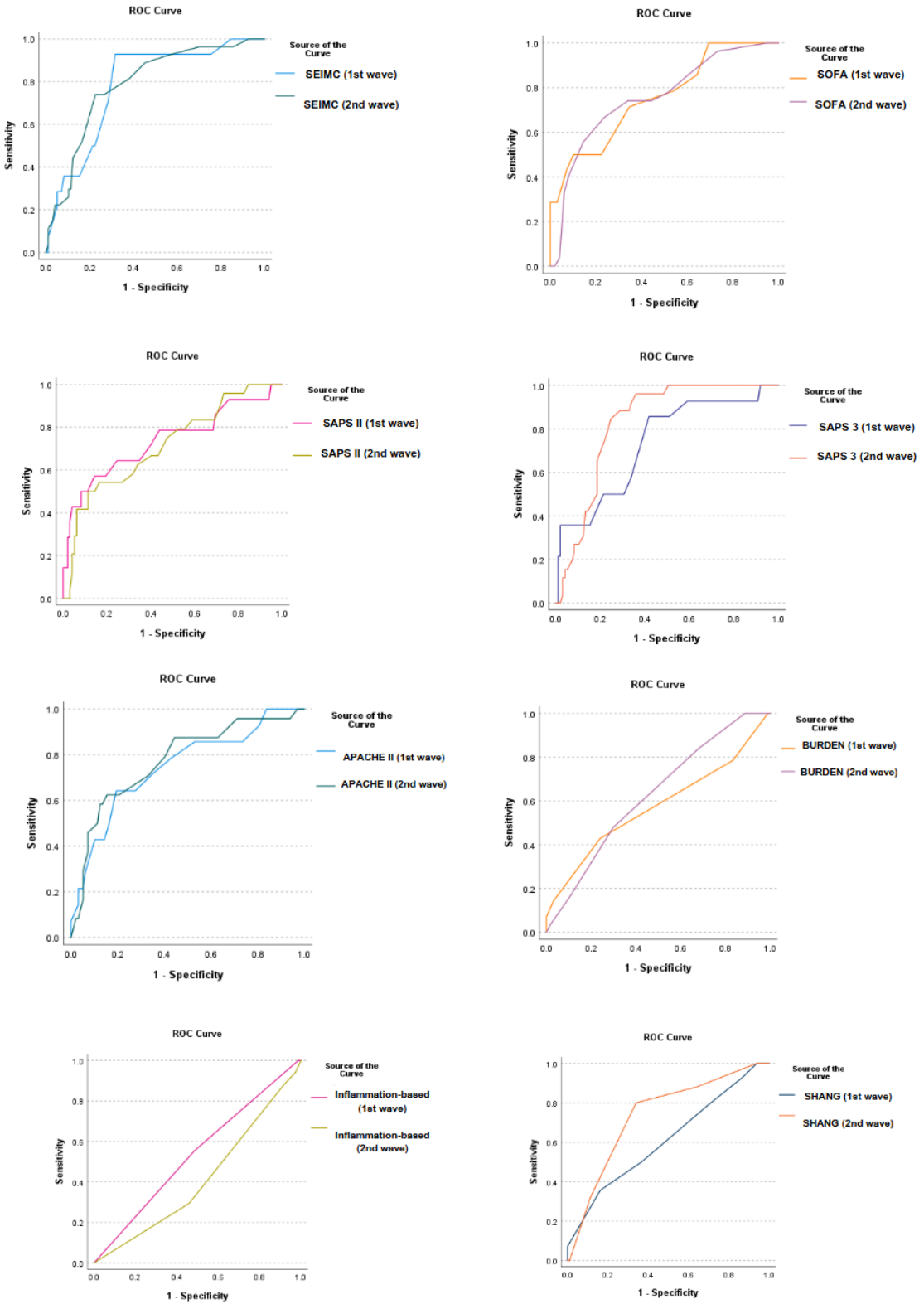

Figure 4.3.3.1 - AUC comparison for each score in first and second waves for early ICU mortality.

4.3.4. Scores comparison between first and second wave for late mortality at ICU

As shown in Table 4.3.4.1 for APACHE II, the AUC is 0.741 in Wave 1 and 0.719 in Wave 2, both values indicating good discriminatory ability, although there is a slight decrease in performance from Wave 1 to Wave 2. The p-value of 0.788 suggests no statistically significant difference in performance between the two waves.

SAPS II shows an AUC of 0.697 in Wave 1 and 0.686 in Wave 2, both of which indicate moderate discriminatory power. The slight decline in AUC from Wave 1 to Wave 2 reflects a minor reduction in predictive accuracy. The p-value of 0.893 indicates no significant difference between the waves.

SAPS 3 performs well with an AUC of 0.832 in Wave 1, but drops to 0.634 in Wave 2, suggesting a notable decrease in its ability to distinguish between outcomes in the second wave. The p-value of 0.009 indicates a statistically significant difference in performance between the two waves.

For SOFA, the AUC is 0.710 in Wave 1 and 0.646 in Wave 2, both values indicating moderate discriminatory power. The slight drop in performance between waves is reflected by the p-value of 0.502, which shows no significant difference in discriminatory ability across the waves.

SEIMC has a high AUC of 0.836 in Wave 1, which drops to 0.641 in Wave 2, indicating a decrease in discriminatory ability. The p-value of 0.008 suggests a statistically significant difference between the two waves, reflecting the change in performance.

SHANG shows an AUC of 0.791 in Wave 1, which drops significantly to 0.576 in Wave 2. This substantial decline in performance is confirmed by the p-value of 0.004, indicating a statistically significant difference between the two waves.

BURDEN has AUC values of 0.586 in Wave 1 and 0.592 in Wave 2, both close to the diagonal of the ROC curve, indicating weak discriminatory ability. The p-value of 0.957 suggests no significant difference in performance between the two waves.

Finally, INFLAMMATION-BASED shows low AUC values of 0.559 in Wave 1 and 0.531 in Wave 2, indicating poor discriminatory power in both waves. The p-value of 0.806 suggests no significant difference between the waves.

In summary, SAPS 3 and SEIMC show the best performance, although both experience a decline in discriminatory ability in Wave 2. APACHE II and SOFA maintain relatively stable performance, while SHANG demonstrates a considerable drop in Wave 2. BURDEN and INFLAMMATION-BASED show weak performance across both waves, with minimal change in their AUC values.

Table 4.3.4.1 - Comparison of severity scores between first and second waves for late mortality outcome.

Scores	Wave	AUC estimate	95% Confidence Interval		p-value
			Lower Bound	Upper Bound	
APACHE II	1	0.741	0.612	0.870	0.788
	2	0.719	0.624	0.814	
SAPS II	1	0.697	0.556	0.839	0.893
	2	0.686	0.587	0.784	
SAPS 3	1	0.832	0.717	0.948	0.009
	2	0.634	0.539	0.729	
SOFA	1	0.710	0.557	0.863	0.502
	2	0.646	0.542	0.751	
SEIMC	1	0.836	0.739	0.932	0.008
	2	0.641	0.537	0.746	
SHANG	1	0.791	0.692	0.890	0.004
	2	0.576	0.467	0.685	
BURDEN	1	0.586	0.410	0.762	0.957
	2	0.592	0.486	0.697	
INFLAMMATION-BASED	1	0.559	0.372	0.746	0.806
	2	0.531	0.412	0.651	

As shown in figure 4.3.4.1, APACHE II demonstrates a stable performance across both waves, with an AUC of 0.741 in Wave 1 and 0.719 in Wave 2. The ROC curve for both waves remains above the diagonal, indicating a strong ability to distinguish between the positive and negative outcomes. However, there is a slight decrease in the steepness of the curve from Wave 1 to Wave 2, which corresponds to a small reduction in discriminatory power.

SAPS II shows a moderate ROC curve performance, with AUCs of 0.697 in Wave 1 and 0.686 in Wave 2. Although both waves indicate reasonable discriminatory ability, the ROC curve is slightly less pronounced than that of APACHE II, and the slight decrease from Wave 1 to Wave 2 suggests a minor reduction in predictive ability.

SAPS 3 has a strong ROC curve in Wave 1 with an AUC of 0.832, but in Wave 2, the curve flattens significantly (AUC of 0.634). This drop suggests a reduced ability to discriminate between outcomes in Wave 2. The curve in Wave 1 indicates a more robust discriminatory ability, but the lower AUC in Wave 2 indicates a weaker predictive power.

For SOFA, the ROC curve remains moderate in both waves, with AUCs of 0.710 and 0.646, respectively. The curve in Wave 1 is steeper, indicating better discriminatory performance, but this steepness diminishes in Wave 2, signalling a reduction in the model's ability to distinguish outcomes.

SEIMC shows a strong ROC curve in Wave 1 with an AUC of 0.836, which indicates good discriminatory power, but it flattens out in Wave 2 (AUC of 0.641), reflecting a decline in predictive accuracy between the two waves. The change in the ROC curve steepness suggests a notable reduction in performance in Wave 2.

SHANG displays a decent ROC curve in Wave 1 (AUC of 0.791), but the curve becomes much flatter in Wave 2 (AUC of 0.576), indicating a substantial loss of discriminatory ability. The ROC curve's decrease in steepness between the two waves reflects a significant decline in its ability to differentiate between outcomes.

BURDEN shows relatively flat ROC curves in both waves, with AUCs of 0.586 in Wave 1 and 0.592 in Wave 2. The curves are close to the diagonal, indicating poor discriminatory ability in both waves, with only a slight improvement in Wave 2. The ROC curves for both waves are weak, reinforcing the model's limited ability to distinguish outcomes.

INFLAMMATION-BASED shows similarly weak ROC curves, with AUCs of 0.559 in Wave 1 and 0.531 in Wave 2. These curves are close to the diagonal, indicating poor performance in both waves, with a slight decrease in Wave 2, which suggests a further reduction in the model's ability to discriminate between outcomes.

In summary, the ROC curves reflect strong discriminatory abilities for APACHE II, SAPS 3, and SEIMC in Wave 1, but with varying declines in performance by Wave 2. SHANG, BURDEN, and INFLAMMATION-BASED consistently show weak ROC curves, indicating poor ability to distinguish between outcomes across both waves.

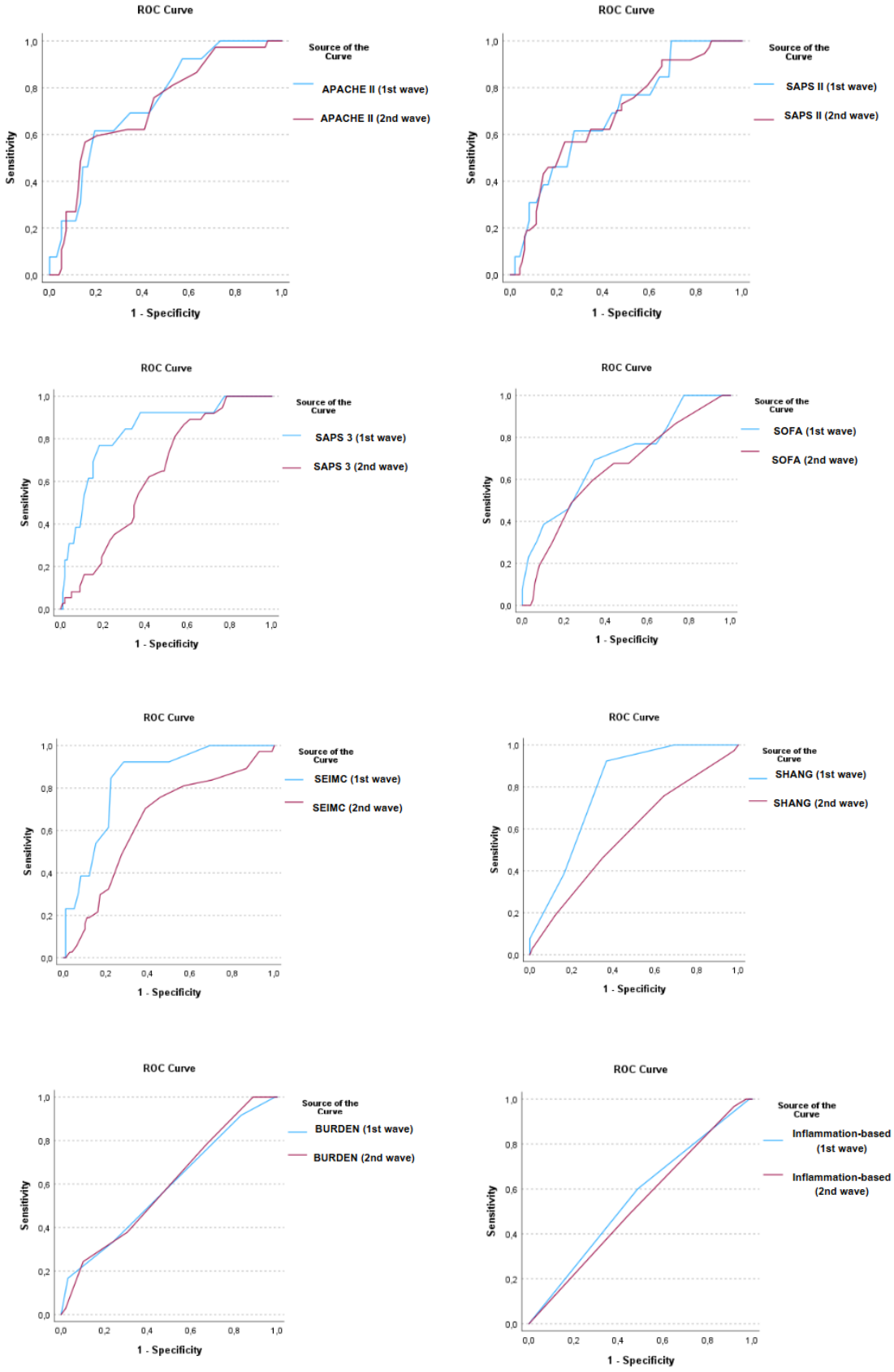

Figure 4.3.4.1 - Scores comparison in first and second waves for late ICU mortality outcome.

5. Conclusion and future perspectives

For COVID-19 first wave data, the analysis included 125 critically ill COVID-19 patients providing valuable insights into the characteristics, outcomes, and predictive performance of various scoring systems in this population. The cohort had a median age of 67 years (P_{25} = 53.50; P_{75}= 76.0), with a notable predominance of males (79.2%). Comorbidities were highly prevalent, affecting 91.2% of patients, underscoring the vulnerability of individuals with pre-existing health conditions to severe COVID-19.

The severity of respiratory involvement was evident, with 80% of patients requiring invasive mechanical ventilation (IMV) and 8% needing extracorporeal membrane oxygenation (ECMO). The outcomes showed significant mortality rates, with 21.6% of patients dying in the ICU and an overall hospital mortality of 28%. The median ICU stay of 9 days (P_{25} = 4.50; P_{75}= 15.0 indicates the prolonged care often required by these patients.

Severity scores demonstrated the critical nature of these cases, with median values of APACHE II at 15.0, SAPS II at 50.0, SAPS 3 at 59.0, and SOFA at 7.0. These scores not only reflect the severity of illness but also showed varying degrees of predictive performance for different mortality outcomes. The comparison of scoring systems revealed that the SEIMC score consistently demonstrated the strongest predictive performance across all mortality outcomes, with AUC values ranging from 0.770 to 0.800 ($p<0.01$) for all outcomes. This suggests that the SEIMC score may be particularly well-suited for risk stratification in COVID-19 patients, in ICU context. SAPS 3 also showed strong performance, especially for hospital mortality (AUC 0.732, p=0.001), ICU mortality (AUC 0.719, p=0.003), and late mortality (AUC 0.797, p=0.002). However, its performance for early mortality was less robust (AUC 0.634, p=0.192). The Shang COVID score performed well for hospital mortality (AUC 0.708, p=0.002), ICU mortality (AUC 0.687, p=0.012), and late mortality (AUC 0.757, p=0.009), but showed weaker performance for early mortality (AUC 0.609, p=0.288). APACHE II demonstrated moderate performance across all mortality types, with AUCs ranging from 0.649 to 0.702, and statistically significant results for hospital mortality, ICU mortality, and late mortality. SOFA showed consistent but moderate performance across all outcomes, with AUC values between 0.642 and 0.664, and statistically significant results for hospital and ICU mortality. In contrast, the Inflammation-based and BURDEN scores generally showed poor predictive ability across all mortality types, with AUCs close to or below 0.6 and non-significant p-values in most cases.

These findings highlight the complex nature of COVID-19 in critically ill patients and the challenges in predicting outcomes. The SEIMC score emerges as a particularly promising tool for risk stratification, while established severity scores like SAPS 3 and APACHE II also demonstrate good overall performance. The results underscore the importance of using appropriate scoring systems for different aspects of mortality risk assessment in COVID-19

patients, which can aid in clinical decision-making and resource allocation in intensive care settings. Future research should focus on validating these findings in larger, diverse cohorts and exploring the potential for combining different scoring systems to improve predictive accuracy.

For COVID-19 second wave data, the analysis included 161 critically ill COVID-19 patients providing valuable insights into the characteristics, outcomes, and predictive performance of various scoring systems in this population. The cohort had a median age of 67 years (P_{25} = 55.0; P_{75}= 76.50), with a notable predominance of males (71.4%). Comorbidities were highly prevalent, affecting 85.7% of patients, underscoring the vulnerability of individuals with pre-existing health conditions to severe COVID-19. The severity of respiratory involvement was evident, with 59% of patients requiring invasive mechanical ventilation (IMV) and 12.4% needing extracorporeal membrane oxygenation (ECMO). The outcomes were particularly concerning, with high mortality rates both in the ICU (39.8%) and overall, in the hospital (47.8%). The median ICU stay of 8 days (P_{25} = 4.0; P_{75}= 14.0), indicates the prolonged care often required by these patients.

Severity scores demonstrated the critical nature of these cases, with median values of APACHE II at 15.0, SAPS II at 48.0, SAPS 3 at 59.50, and SOFA at 6.0. These scores not only reflect the severity of illness but also showed varying degrees of predictive performance for different mortality outcomes.

The comparison of scoring systems revealed that SAPS II, APACHE II, and SAPS 3 consistently demonstrated strong predictive performance across different mortality outcomes. SAPS II showed the highest discriminative ability for hospital mortality (AUC 0.756, $p<0.001$), while SAPS 3 performed best for early mortality (AUC 0.791, $p<0.001$). APACHE II demonstrated the most consistent performance across all mortality types, with AUCs ranging from 0.704 to 0.743 (all $p<0.001$).

Interestingly, the SEIMC score showed particularly strong performance for early mortality prediction (AUC 0.813, $p<0.001$), outperforming other scores in this category. The Shang COVID score, while performing moderately for most outcomes, showed strong predictive ability for early mortality (AUC 0.772, $p=0.001$). In contrast, the Inflammation-based and BURDEN scores generally showed poor predictive ability across all mortality types, with AUCs close to or below 0.5 and non-significant p-values in most cases.

These findings highlight the complex nature of COVID-19 in critically ill patients and the challenges in predicting outcomes. While established severity scores like SAPS II, APACHE II, and SAPS 3 demonstrate good overall performance, newer COVID-specific scores like SEIMC and Shang COVID show promise, particularly for early mortality prediction. The results underscore the importance of using appropriate scoring systems for different aspects of

mortality risk assessment in COVID-19 patients, which can aid in clinical decision-making and resource allocation in intensive care settings.

The analysis of outcomes for different scoring systems across multiple waves reveals a range of results, with most showing little to no significant change. For hospital mortality, the APACHE II, SAPS II, SAPS 3, SOFA, SEIMC, SHANG, BURDEN, and INFLAMMATION-BASED scores exhibited either small increases or decreases in AUC values, but none showed significant changes based on the p-values, which ranged from 0.146 to 0.656. The SEIMC score, despite showing the highest AUC in wave 1, saw a decrease in wave 2, but this was not statistically significant (p-value of 0.146). The INFLAMMATION-BASED score had a decline in AUC, and its p-value also suggested no significant difference.

For ICU mortality, the APACHE II and SAPS II scores showed stable AUC values, with no significant differences between waves, as indicated by p-values of 0.973 and 0.800, respectively. The SAPS 3 score had a decline in AUC from 0.781 to 0.719, but the change was not statistically significant (p-value of 0.332). The SEIMC score showed a slight decline in AUC from wave 1 to wave 2 (from 0.808 to 0.705), with a borderline significant p-value of 0.087. Other scores like SHANG, BURDEN, and INFLAMMATION-BASED also showed no significant changes, with p-values ranging from 0.456 to 0.748.

In terms of early mortality, the APACHE II score showed a slight improvement in AUC from 0.747 to 0.775, but the change was not statistically significant (p-value of 0.762). The SAPS II score decreased slightly from 0.737 to 0.716, with a p-value of 0.843, indicating no significant change. The SAPS 3 score showed a notable improvement, with the AUC rising from 0.733 to 0.828, but the p-value of 0.240 suggests no significant difference. The SOFA score remained nearly unchanged, while the SEIMC score showed a marginal increase from 0.783 to 0.787, with a p-value of 0.964, indicating no significant improvement. The INFLAMMATION-BASED score saw a decline in AUC, suggesting no improvement over time.

For late mortality, the SAPS 3, SEIMC, and SHANG scores showed a significant decline in AUC. SAPS 3 had a significant drop in performance from 0.832 to 0.634, with a p-value of 0.009. The SEIMC score also showed a significant decline from 0.836 to 0.641 (p-value of 0.008), and the SHANG score dropped significantly from 0.791 to 0.576 (p-value of 0.004). These results indicate that the predictive performance of these models worsened over time. In contrast, the APACHE II, SAPS II, SOFA, BURDEN, and INFLAMMATION-BASED scores showed little change, with p-values suggesting no significant differences in predictive ability for late mortality.

In summary, while a few scoring systems, particularly SAPS 3, SEIMC, and SHANG, demonstrated some changes in predictive accuracy over time, most models showed either marginal or no significant changes across the different mortality outcomes. This indicates that,

for most models, predictive performance remained stable, with few significant improvements or deteriorations observed across the waves.

About COVID-19-specific scores, it is important to note that these scores were validated with populations exposed to the virus during the first wave of the pandemic. The first wave was marked by the emergence of the SARS-CoV-2 virus, and the populations involved had different demographic, clinical, and genetic characteristics. The loss of efficacy of these scores in the second wave can possibly be explained by changes in the population and disease characteristics. For COVID-19 specific scores, the populations with which they were validated are equivalent to the first wave, and this may explain their loss of discriminatory capacity in the second wave. Additionally, the evolution of the virus and the introduction of new variants could have altered the disease's progression, making triage tools less accurate.

Future perspectives in the use of severity scores for COVID-19 prognosis should focus on enhancing the accuracy and applicability of these tools across different waves of the pandemic, as well as in diverse patient populations. While traditional scores such as APACHE II, SAPS II, and SAPS 3 have demonstrated reliable predictive power, their performance may be influenced by changes in patient demographics, treatment protocols, and healthcare resources over time. Therefore, continuous validation and recalibration of these scores are necessary to maintain their relevance and ensure they reflect the evolving nature of the disease and its management.

Additionally, scores specifically designed for COVID-19, such as SEIMC, offer promising insights. Future studies could explore the integration of emerging biomarkers, imaging techniques, and machine learning models to further enhance their predictive accuracy. Incorporating real-time data from advanced monitoring systems, such as continuous vital sign monitoring or artificial intelligence algorithms, may help refine these tools, allowing for more precise and dynamic prognostic assessments.

Another key area for future research is the development of scores that can account for individual patient responses to treatment. Personalized medicine, which takes into consideration genetic, immunological, and environmental factors, could pave the way for more tailored approaches to patient care. Scores that integrate these factors alongside traditional clinical data may improve risk stratification and allow for better resource allocation, particularly in resource-constrained settings.

Moreover, the integration of these severity scores into electronic health records (EHRs) and decision support systems could streamline their use in clinical practice. By providing real-time prognostic assessments and facilitating early interventions, these systems could enhance patient outcomes and optimize the use of ICU and hospital resources during future pandemics or crises.

Finally, interdisciplinary collaborations between clinicians, data scientists, and public health experts will be essential in refining and validating these scores, ensuring they are adaptable and capable of meeting the challenges posed by future global health emergencies.

References

1. Amezcua-Guerra, L. M., Audelo, K., Guzmán, J., Santiago, D., González-Flores, J., García-Ávila, C., Torres, Z., Baranda-Tovar, F., Tavera-Alonso, C., Sandoval, J., & González-Pacheco, H. (2021). A simple and readily available inflammation-based risk scoring system on admission predicts the need for mechanical ventilation in patients with COVID-19. Inflammation Research, 70(6), 731–742. https://doi.org/10.1007/s00011-021-01466-x

2. Berenguer, J., Borobia, A. M., Ryan, P., Rodríguez-Baño, J., Bellón, J. M., Jarrín, I., Carratalà, J., Pachón, J., Carcas, A. J., Yllescas, M., & Arribas, J. R. (2021). Development and validation of a prediction model for 30-day mortality in hospitalised patients with COVID-19: The COVID-19 SEIMC score. Thorax, 76(9), 920–929. https://doi.org/10.1136/thoraxjnl-2020-216001

3. Chaw, S. M., Tai, J. H., Chen, S. L., Hsieh, C. H., Chang, S. Y., Yeh, S. H., Yang, W. S., Chen, P. J., & Wang, H. Y. (2020). The origin and underlying driving forces of the SARS-CoV-2 outbreak. Journal of Biomedical Science, 27(1), 1–12. https://doi.org/10.1186/s12929-020-00665-8

4. Chung, H. P., Tang, Y. H., Chen, C. Y., Chen, C. H., Chang, W. K., Kuo, K. C., Chen, Y. T., Wu, J. C., Lin, C. Y., & Wang, C. J. (2023). Outcome prediction in hospitalized COVID-19 patients: Comparison of the performance of five severity scores. Frontiers in Medicine, 10(February), 1–9. https://doi.org/10.3389/fmed.2023.1121465

5. Constipação, gripe e COVID-19 | Hospital da Luz. (2020). Retrieved September 30, 2024, from https://www.hospitaldaluz.pt/pt/saude-e-bem-estar/constipacao-gripe covid-19

6. COVID-19. (2023). Retrieved September 30, 2024, from https://www.sns24.gov.pt/tema/doencas-infecciosas/covid-19/

7. DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics, 44(3), 837. https://doi.org/10.2307/2531595

8. DGS. (2022). Relátorio de Situação - 01/11/2022 - 07/11/2022. 21–24.

9. Direção Geral de Saúde. (2020). Relatório de Situação N.º 16 - 18/03/2020. Direção Geral de Saúde, 2020, 1–2. https://www.dgs.pt/em-destaque/relatorio-de situacao-n-016-18032020-pdf.aspx

10. Early COVID-19 symptoms differ among age groups, research finds | King's College London. (2021). Retrieved September 8, 2024, from https://www.kcl.ac.uk/news/early-covid-19-symptoms-differ-among-age-groups

11. Fim do estado de alerta - XXIII Governo - República Portuguesa. (2022). Retrieved September 29, 2024, from https://www.portugal.gov.pt/pt/gc23/comunicacao/noticia?i=fim-do-estado-de-alerta

12. Hao, Y. J., Wang, Y. L., Wang, M. Y., Zhou, L., Shi, J. Y., Cao, J. M., & Wang, D. P. (2022). The origins of COVID-19 pandemic: A brief overview. Transboundary and Emerging Diseases, 69(6), 3181–3197. https://doi.org/10.1111/tbed.14732

13. Imanieh, M. H., Amirzadehfard, F., Zoghi, S., Sehatpour, F., Jafari, P., Hassanipour, H., Feili, M., Mollaie, M., Bostanian, P., Mehrabi, S., Dashtianeh, R., & Feili, A. (2023). A novel scoring system for early assessment of the risk of the COVID 19-associated mortality in hospitalized patients: COVID-19 BURDEN. European Journal of Medical Research, 28(1), 1–7. https://doi.org/10.1186/s40001-022-00908-4

14. INFARMED - Autoridade Nacional do Medicamento e Produtos de Saúde, I.P. (2022). Terapêuticas farmacológicas disponíveis para a COVID-19.

15. Martin, J., Gaudet-Blavignac, C., Lovis, C., Stirnemann, J., Grosgurin, O., Leidi, A., Gayet-Ageron, A., Iten, A., Carballo, S., Reny, J. L., Darbellay-Fahroumand, P., Berner, A., & Marti, C. (2022). Comparison of prognostic scores for inpatients with COVID-19: A retrospective monocentric cohort study. BMJ Open Respiratory Research, 9(1), 1–9. https://doi.org/10.1136/bmjresp-2022-001340

16. Moreno, R. P., Metnitz, P. G. H., Almeida, E., Jordan, B., Bauer, P., Campos, R. A., Iapichino, G., Edbrooke, D., Capuzzo, M., & Gall, J. L. (2005). SAPS 3—From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. Intensive Care Medicine, 31(10), 1345–1355. https://doi.org/10.1007/s00134-005-2763-5

17. Nahm, F. S. (2022). Receiver operating characteristic curve : overview and practical use for clinicians.

18. NCSS, LLC. (2021). Comparing two ROC curves – paired design. In NCSS Statistical Software (pp. 547–1).

19. Nogueira, P. J., de Araújo Nobre, M., Elias, C., Feteira-Santos, R., Martinho, A. C. V., Camarinha, C., Bacelar-Nicolau, L., Costa, A. S., Furtado, C., Morais, L., Rachadell, J., Pinto, M. P., Pinto, F., & Carneiro, A. V. (2022). Multimorbidity Profile of COVID-19 Deaths in Portugal during 2020. Journal of Clinical Medicine, 11(7), 1–19. https://doi.org/10.3390/jcm11071898

20. NÚMERO DE NOVOS CASOS E ÓBITOS POR DIA - Covid 19. (2024). Retrieved September 29, 2024, from https://covid19.min-saude.pt/numero-de-novos casos-e-obitos-por-dia/

21. Rapsang, A. G., & Shyam, D. C. (2014). Scoring systems in the intensive care unit: A compendium. Indian Journal of Critical Care Medicine, 18(4), 220–228. https://doi.org/10.4103/0972-5229.130573

22. Sakr, Y., Krauss, C., Amaral, A. C. K. B., Réa-Neto, A., Specht, M., Reinhart, K., & Marx, G. (2008). Comparison of the performance of SAPS II, SAPS 3, APACHE II, and their customized prognostic models in a surgical intensive care unit. British Journal of Anaesthesia, 101(6), 798–803. https://doi.org/10.1093/bja/aen291

23. Shang, Y., Liu, T., Wei, Y., Li, J., Shao, L., Liu, M., Zhang, Y., Zhao, Z., Xu, H., Peng, Z., Zhou, F., & Wang, X. (2020). Scoring systems for predicting mortality for severe patients with COVID-19. EClinicalMedicine, 24 (December 2019), 100426. https://doi.org/10.1016/j.eclinm.2020.100426

24. SICO – SPMS. (2013). Retrieved September 30, 2024, from https://www.spms.min-saude.pt/2013/10/sico/

25. Siddiqui, S. S., Patnaik, R., & Kulkarni, A. P. (2022). General Severity of Illness Scoring Systems and COVID-19 Mortality Predictions: Is "Old Still Gold?" Indian Journal of Critical Care Medicine, 26(4), 416–418. https://doi.org/10.5005/jp-journals-10071- 24197

26. Unim, B., Palmieri, · Luigi, Cinzia, ·, Noce, L., Brusaferro, S., & Graziano Onder, ·. (2021). Prevalence of COVID-19-related symptoms by age group. 33, 1145–1147. https://doi.org/10.1007/s40520-021-01809-y

27. Von Rekowski, C. D. P. (2022). Development of Predictive Models for COVID-19 Prognosis based on Patients' Demographic and Clinical Data. Instituto Superior de Engenharia de Lisboa