Pontifícia Universidade Católica do Rio de Janeiro

Leonardo dos Santos Lourenço Bastos

Analysis of Performance in Intensive Care Units

Dissertação de Mestrado

Dissertation presented to the Programa de Pós-Graduação em Engenharia de Produção of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Engenharia de Produção

Advisor: Prof. Silvio Hamacher
Co-advisor: Prof. Fernando Augusto Bozza

Rio de Janeiro
February 2018

Leonardo dos Santos Lourenço Bastos

Analysis of Performance in Intensive Care Units

Dissertation presented to the Programa de Pós-Graduação em Engenharia de Produção of PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Engenharia de Produção. Approved by the undersigned Examination Committee.

Prof. Silvio Hamacher
Advisor
Departamento de Engenharia Industrial - PUC-Rio

Prof. Fernando Augusto Bozza
Co-Advisor
Fundação Oswaldo Cruz - FIOCRUZ

Prof. Fernando Godinho Zampieri
Universidade de São Paulo - USP

Prof. Julia Lima Fleck
Departamento de Engenharia Industrial - PUC-Rio

Prof. Márcio da Silveira Carvalho
Vice Dean of Graduate Studies
Centro Técnico Científico - PUC-Rio

Rio de Janeiro, February 22 th, 2018

Leonardo dos Santos Lourenço Bastos

Graduated in Production/Industrial Engineer at the Universidade do Estado do Pará – UEPA in 2015. Participated in the Núcleo Integrado de Logística e Operações - NILO and the Conceito em Engenharia de Produção Laboratory – CONCepT. Has experience in Logistics & Process Management, using optimization and simulation tools as well as statistical data analysis process improvement.

Bibliographic data

Bastos, Leonardo dos Santos Lourenço

Analysis of performance in intensive care units / Leonardo dos Santos Lourenço Bastos ; advisor: Silvio Hamacher ; co-advisor: Fernando Augusto Bozza. – 2018.
91 f. : il. color. ; 30 cm

Dissertação (mestrado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Industrial, 2018.
Inclui bibliografia

1. Engenharia Industrial – Teses. 2. Unidade de terapia intensiva. 3. Análise de desempenho. 4. Matriz de eficiência. 5. Mortalidade. 6. Perfil de risco. I. Hamacher, Silvio. II. Bozza, Fernando Augusto. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Industrial. IV. Título.

CDD: 658.5

To my parents, Carlos and Mônica, my sister, Isabela, and
my nephew, Carlos Eduardo.

Acknowledgements

To God for all the strength, bless and guidance that could assist me towards my objectives.

To my parents, Carlos and Monica, my sister Isabela and my nephew Carlos Eduardo, my grandmother Elizete, my aunts, uncles and cousins, for the love and support always present in my journeys.

To Verissa, for all the love, support, friendship, and motivation given to me towards my dreams and objectives, even in the hardest of the moments.

To my longtime friends, for the constant partnership and support. Despite the distance we are always close somehow and this really motivates me.

To my new friends from this recent journey, for introducing me to new universes and sharing your precious lessons. We have grown together, and I wish you all the best.

To my advisor, Silvio Hamacher, for all the learning, lessons, brainstorms, and opportunities given, which certainly assisted the development of this and other projects.

To Fernando Bozza, for the ideas, insights, the opportunity of performing those analysis and the lessons on the healthcare world.

To all the Professors, colleagues and employees at DEI, for providing the assistance and the necessary resources that conveyed the development of this dissertation and myself
.
To the researchers and colleagues at IDOR, and HCor, Fernando Zampieri, Alexandre Biasi and Marcio Soares, for assisting in the development and understanding of this large amount of information and providing grateful insights.

To CNPq and PUC-Rio, for providing adequate resources and the opportunity to participate in this great university.

To all of those who contributed somehow for the fulfillment of this work.

Abstract

Bastos, Leonardo dos Santos Lourenço; Hamacher, Silvio (Advisor); Bozza, Fernando Augusto (Co-Advisor). **Analysis of Performance in Intensive Care Units**. Rio de Janeiro, 2018. 91p. Dissertação de Mestrado - Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro.

Intensive Care Unit (ICU) is an important department within a hospital since it deals mostly with complex cases and it generates the highest amount of costs, thus requiring adequate control on its care treatments. Nonconformities such as poor communication and treatment errors are commonly responsible for a bad performance in ICUs. However, evaluating the performance of an ICU is not an easy task and there are no gold-standard indicators. The most common metrics are the Standardized Mortality Ratio (SMR) and the Standardized Resource Use (SRU), which measure mortality and resource utilization, respectively. Hence, this study aims to analyze different ICUs in terms of mortality, resource use, and institutional factors, combining the methods Efficiency Chart, Rankability and Risk Profile. The analysis was performed considering a total of 12,100 patients in 116 ICUs provided by a clinical trial study. As results, it was verified that most ICUs were from hospitals with public administration (47.41%), which had significantly high lethality rate compared to private hospitals. Four different clustering approaches were tested, which identified similar case-mixes between the best and lower performance groups of ICUs, and a high variability in expected risks for low severity patients. Using a resampling approach, it was evidenced that the mortality indicator varies strongly on low-risk groups of patients, while high-risk patients had a smaller range of SMR values, which may lead to biased conclusions when comparing ICUs with similar mortality and different case-mixes.

Keywords

Intensive care units; benchmarking, efficiency matrix, risk profiles, mortality.

Resumo

Bastos, Leonardo dos Santos Lourenço; Hamacher, Silvio (Orientador); Bozza, Fernando Augusto (Co-orientador). **Análise de Performance em Unidades de Terapia Intensiva**. Rio de Janeiro, 2018. 91p. Dissertação de Mestrado - Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro.

A Unidade de Terapia Intensiva (UTI) é um departamento importante dentro do Hospital visto que lida majoritariamente com casos de alta complexidade e gera elevados custos administrativos, o que requer um controle adequado de seus processos. Inconformidades tais como erros em atividades de tratamento e falta de comunicação entre os funcionários são comumente responsáveis pelo baixo desempenho de UTIs e devem ser ajustados para reduzir possíveis danos ao tratamento do paciente. Para avaliar a eficiência de uma UTI, a literatura propõe que sejam estabelecidas métricas que considerem quatro perspectivas - médica ou clínica, econômica, social e institucional – que oferecem uma visão abrangente das atividades (administrativas ou de tratamento) dentro da unidade e seus impactos no pós-tratamento. Entretanto, a avaliação de desempenho em uma UTI não é uma tarefa simples, pois há diversas variáveis a serem consideradas e que podem ser potenciais causas de um mau desempenho. Além disso, não há uma métrica ou indicador "padrão-ouro" que consegue reter de forma adequadas as informações, sendo que diversas perspectivas devem ser consideradas. Os indicadores mais comuns são A Taxa de Mortalidade Padronizada (*Standardized Mortality Ratio*, SMR) e o Taxa de Uso de Rescursos Padronizada (*Standardized Resource Use*, SRU), que contabilizam desfechos de mortalidade (clínicos) e de uso de recursos (econômicos), junto de metodologias propostas para viabilizar a comparação entre diferentes UTIs, identificar de grupos de desempenho e analisar os riscos de mortalidade dos pacientes dentro da unidade, tais como os conceitos de *Rankability* e Perfis de Risco (*Risk Profiles*). Além disso, é necessário definir corretamente os desfechos a serem contabilizados em indicadores. Nesse contexto, recomenda-se a combinação de diferentes indicadores e metodologias de forma a complementar e elevar a confiabilidade da análise de desempenho e *benchmarking*. Com isso, este estudo tem como objetivo analisar um conjunto de UTIs em termos de desempenho quanto à mortalidade e uso de recursos, associando-os com as características das unidades e seus fatores institucionais, para identificar possíveis correlações. A análise foi feita em uma amostra composta por 12.100 pacientes que foram hospitalizados em 116 UTIs, considerando um desfecho em até 60 dias de interação. Este estudo teve como contribuição a combinação de diferentes técnicas e indicadores, e uma discussão a respeito da variabilidade do SMR em comparação à metodologia tradicional. Para este propósito, combinou-se as técnicas da Matriz de Eficiência, Rankability – índice de confiabilidade de um indicador de desfecho, e Perfis de Risco, de forma a obter e avaliar o desempenho de grupos de UTIs. Como resultados, verificou-se que UTIs cuja administração é de domínio Público e que destinam a maioria dos seus leitos ao Sistema Único de Saúde (SUS) brasileiro tiveram mortalidade significativamente alta em relação àquelas de domininío privado (p-valor < 0.05). Além disso, realizou-se um agrupamento das UTIs utilizando quatro diferentes técnicas de clusterização de forma a garantir a máxima confiabilidade do indicador para comparação (*Rankability*), o que resultou na presença de clusters extremos contendo uma UTI cada, sendo elas a de maior e a de menor SMR, apesar de ambas apresentarem o mesmo conjunto de severidades. Para cada grupo, estimou-se o seu perfil de risco, e verificou-se que pacientes com menor gravidade apresentaram maior variabilidade nos riscos de morte, sendo estes maiores nos grupos com alto SMR e menores em grupos de menor mortalidade, sendo que a dispersão tendeu a ser menor quanto menor for o risco, o que poderia influenciar diretamente no cálculo do SMR. Com isso, por meio de equações matemáticas e simulação por meio de reamostragem, verificou-se que o SMR possui uma limitação em sua escala, que depende diretamente do espectro de gravidade dos pacientes em cada UTI ou grupo de desempenho analisado. O SMR possui maior variabilidade para grupos de gravidade de baixo risco enquanto que o alto risco diminui esse intervalo, demonstrando um possível viés, o que pode resultar em conclusões enviesadas ao comparar UTIs com mesmo valor de SMR, porém com diferentes *case-mixes*.

Palavras-chave

Unidade de terapia intensiva, análise de performance, matriz de eficiência, perfil de risco, mortalidade.

Table of Contents

List of Figures

List of Tables

Pass on what you have learned. Strength, mastery. But weakness, folly, failure also. Yes, failure most of all. The greatest teacher, failure is. […], we are what they grow beyond. That is the true burden of all masters.

Master Yoda, *Star Wars Episode VIII – The Last Jedi*

1
Introduction

Organizations have studied several approaches for monitoring their processes to improve their quality and reduce costs. Usually, they create performance indicators, with the objective of obtaining correct information and references from their processes to assist planning decisions. Besides cost, other variables have been accounted, such as customer satisfaction, which may vary depending on the product or service to be performed.

In this context, performance indicators have been used not only by manufacturing companies, but also by service industries. In the latter, Healthcare services have applied those metrics with the objective to control and guarantee adequate treatment levels and care for patients. As Healthcare systems deal directly with human lives, it is important for their process to provide the most adequate care, considering health treatment, ethical and psychosocial effects, as well as reducing costs and improving efficiency.

Within a Hospital, activities are executed depending on the specialty needed or the case-mix, severity of patients and the management objectives. Among healthcare services, the Intensive Care, represented by the Intensive Care Units (ICU), is an important department since it deals with complex cases and results in larger costs compared to other departments (Garland, 2005; Ray et al., 2009). According to Ray et al. (2009), despite the "capital intensive" care, ICUs save lives as well as contribute to the comprehension regarding the course of a disease, assisting future actions.

Therefore, an ICU must operate effectively and efficiently in its treatments, since it must ensure the adequate care with the correct treatment as well as comprehend the management's objectives and planning for its sustainability. Comparison between ICUs as benchmarks can provide better targets from improvements, in areas such as mortality, safety, processes of care, economic outcomes and patient satisfaction (Salluh et al., 2017; Woodhouse et al., 2009). Therefore, it comprehends a major component of healthcare systems, evidences have shown problems within the ICU and great efforts to improve its performance (Garland, 2005).

According to Garland (2005), common problems found in ICUs comprehend nonadherence of processes to standards, poor communications in staff, and errors in treatment, which are related to poor outcomes, low efficiency, and dissatisfaction of patients. Lone et al. (2016) found that the excess in long-term ICU mortality and hospital costs for saving patients may result from the complex interplay among illness factors and hospital organization.

The analysis of ICU performance is not an easy task, since its outcomes are not easily accountable, depending on the nature of the ICU, the country (region), and case-mix (Garland, 2005; Ray et al., 2009). Hence, appropriate performance indicators must be developed to be representative of the processes, especially the use of standardized parameters, as it reduces variability in data and provide comparable information (Brown et al., 2014; Ray et al., 2009). Those performance or quality indicators are useful screening tools for evaluating and identifying potential improvements in healthcare (Brown et al., 2014).

Currently, mortality is the primary outcome verified to evaluate an ICU and the main indicator associated with it in literature is the Standardized Mortality Ratio (SMR), which relates observed to expected mortality outcome. Although the use of two parameters can provide a good overall view of the ICU performance throughout time or comparing to similar units within a network, more indicators should be accounted for when possible.

Occasionally, the analysis of resource utilization has been performed using outcomes related to costs, patient Length-Of-Stay (LOS), or workload scores such as the Therapeutic Intervention Scoring System (TISS) or the Nine Equivalents of Nursing Manpower (NEMS). Rothen et al. (2007) has proposed the Standardized Resource Use (SRU), and indicator similar in scale with the SMR, using LOS as outcome variable for calculation.

The use of those standardized indicators requires a reference value, usually estimated by a prediction model or, when accessible, a reference population, for national or international comparison (Ray et al., 2009). Hence, literature provides a wide set of models that use different variables to predict mortality, using severity scores as predictors, such as SAPS-3, APACHE or MPM, and resource use, considering LOS as a representation of treatment efforts (Rothen et al., 2007; Rothen & Takala, 2008; Salluh et al., 2017)

SMR has been considered one of the best approaches to evaluate mortality by the simplicity to calculate, although it may present bias depending on predicted risks from severity scores (Siegel et al., 2015). One limitation is that the straightforward (direct) comparison of SMR values and the classification using league tables may not provide reliability in comparing "best" and "worst" ICU performance (Verburg et al., 2016). A second limitation has been that SMR is presented as one-value representation of the case-mix severity scores of an ICU, which is not always true, since the unit may perform differently for distinct severity scores (Moreno et al., 2010).

Verburg et al. (2016) have first applied the concept of Rankability and provided a methodology to cluster similar ICUs, and Moreno et al. (2010) proposed the Risk Profiles concept, to evaluate the risk ratio over the severity span of an ICU, instead of relying only in the SMR value. Kramer (2016) claims that clustering ICUs instead of using league tables is a good approach since it deals with the sample size problems of different ranks and analyzing the performance between groups and within each group can reveal the reasons for "better" or "worse" performance.

Conversely, SRU has not been widely used in benchmark studies (Rothen et al., 2007; Salluh et al., 2017), even though the LOS variable is the most used for efficiency measurement by the facility compared to other variables, as for example cost. Brown et al. (2014) analyzed ten different indicators measuring outcomes such as mortality, ICU readmissions, and LOS, and concluded that there is not a gold standard indicator for performance analysis. Salluh et al. (2017) concluded that advances in big data and machine learning tools can provide in the future better approaches for those indicators, ensuring lower bias for analysis.

One can notice that the analysis of ICU performance can be performed considering different types of outcomes. Mortality has been the main outcome to be analyzed, however SMR presents significant variability depending on the expected mortality value from the reference population, while resource use has been related mainly to the length of stay of a patient. The analysis of performance on different datasets and the relation between other variables provide a good contribution to this field of study, furthermore there is room for new proposals of robust indicators that can provide benchmarking and comparison between different ICUs.

Hence, the main objective of this research is to evaluate the performance of an ICU network using the indicators of mortality and resource utilization. For this purpose, we considered the efficiency matrix approach, the Rankability and the Risk Profile Management, to evaluate and compare the performance among different groups of ICUs. We performed the analysis on a database of Brazilian ICUs, presented in Cavalcanti et al. (2016), which offered information about patients (characteristics, comorbidities and treatment), and the ICUs (infrastructure, organization and resources). The complementary objectives of this research are:

- Perform a literature review to identify the main concepts, indicators the outcome variables used in the analysis of performance of ICUS;
- Analyze the fitness of the severity scores equation (SAPS3) to the dataset using Calibration Belts;
- Define the best technique of clustering ICUs using the Rankability methodology applied to the mortality indicator;
- Evaluate the Risk Profile behavior of each cluster regarding to their case-mixes;
- Combine institutional variables to the performance indicators to identify possible patterns and relations;

This research contributes primarily with the combination of two robust techniques, Rankability and Risk Profiles, that allow better comparisons between units, reducing the bias of the mortality outcome indicator and it extends the analysis using comparisons of other outcome variables to evidences from the Risk Profiles. Furthermore, the application of those techniques to a Brazilian dataset of ICUs is a novel approach for performance analysis in the country. Finally, this research adds to the current ongoing study field regarding performance analysis in Intensive Care, especially on the analysis of the current performance indicators, which can also contribute to the overall Healthcare systems management.

Following this introductory section, this research is divided as follows: Section 2 describes the main status and contributions in literature on ICU performance analysis; Section 3 comprises the research steps and tools used for evaluating the database; in Section 4 we present the descriptive analysis of the database regarding patients and ICUs and in Section 5 there are the main results on the performance evaluation and comparisons; Finally, Section 6 states the conclusion and final considerations followed by the References.

2
Performance Analysis and Benchmarking in ICUs

Intensive Care Units (ICUs) have a significant participation in the care and the costs of treatment of severe patients admitted in the hospital (Garland, 2005). Moreover, the assessment of performance within those units is not an easy task, as ICU patients typically present variability in severity of disease, additional risk factors and, therefore, risk of mortality (Rothen & Takala, 2008). Garland (2005) states one should evaluate different parameters in the ICU performance dimension and use performance indicators that are primarily relevant for the unit. Therefore, the different variables regarding ICU processes such as the staff, the infrastructure, case-mix, and patient's satisfaction should be included in the planning decisions.

The most used approach in the ICU performance analysis has been the evaluation of process and outcomes with quality indicators (Rothen & Takala, 2008). According to Garland (2005), performance indicators should be relevant to the patient, society and the hospital. For this purpose, different outcomes have been studied to incorporate and compare ICUs, individually or in group. In this context, Garland (2005) proposes four domains of outcomes to measure ICU performance: medical, economic, psychosocial, and institutional as described in Table 1.

Table 1 - Domains and Measurement of Outcomes. Source: Garland (2005)

Outcomes	**Measures**
Medical	ICU, hospital, and long-term survival-rates; complication rates related to care; medical errors; and adequacy of symptom control
Economic	ICU, hospital, and post-treatment use of resources; and cost-effectiveness of care
Psychosocial and Ethical	Long-term quality of life among survivors; patient and family satisfaction; concordance of expected and actual end-of-life decisions
Institutional	Staff satisfaction and turnover rate; effectiveness of ICU bed utilization; staff satisfaction in the hospital with the care and services supplied by the ICU; and efficiency of processes/procedures/functions involved in ICU care

Although some of those variables are easy to compute, data must be reliable. Mortality and LOS, for instance, are very sensitive to the case-mix of certain hospital, as well as the number of readmissions of an ICU is limited measured within a certain period. Salluh et al. (2017) describe the advantages and limitations of using those variables for measuring performance, as shown in Table 2.

Table 2 – Advantages and Limitations of using certain outcomes. Source: (Salluh et al., 2017)

Domain/Measure	**Advantages**	**Limitations**
Outcomes		
Mortality	Easy to measure, clinically relevant	It must be risk-adjusted with well-calibrated scores; it is sensitive to case-mix.
Length of Stay	Easy to measure, clinically relevant, proxy of efficiency	Affected by structure, can be artificially lowered by transfers
ICU Readmissions	Easy to measure, clinically relevant, indirect marker of clinical process inside and outside ICU. Evaluations of unplanned and early ICU readmissions are preferable as they reflect quality and safety	Affected by structure (e.g., step-down units), artificially lowered by transfers and end of life care policies
ICU Acquired Complications	Indicators of quality of care	Affected by case-mix, frequently under-reported, applied definition may vary
Patient-reported Outcomes	Post-ICU vital status and quality of life	Under-reporting, low-adherence, need for specialized platforms
Processes of Care		
Adherence to best practices	Reliable surrogate of best practices, extensive evidence-based medicine literature to support, can be used for audit-feedback purposes	Level of evidence varies according to the measures, effect on outcomes is variable, frequently under-reported, tricky to measure at bedside, frequently requires specialized monitoring system
ICU and Hospital Organization/Structure		
Staffing Patterns	Potentially associated with outcomes, easy to measure	Should be adjusted by risk and workload
ICU Structure	Can be measured within countries where there are national requirements to provide intensive care; Can allow stratification of levels of care that can be provided by the ICU	Wide variation in national standards as well as in the definition of an ICU bed

Rothen et al. (2007) assert that there is some variability in resource use, such as organizational characteristics, type and size of ICU, and physician staff, which are not widely studied. Furthermore, the evaluation of performance should not rely only on indicators, since they may present bias due to data collection, and process improvement should be the focus while comparing to benchmarks (Rothen & Takala, 2008). Considering medical and economic outcomes, two main indicators discussed in literature are mortality and resource use, which are described in what follows.

2.1
Mortality Indicators

Mortality is the main primary outcome measure for evaluating quality of care used in literature (Siegel et al., 2015). The main approach is to compare the observed number of deaths within the sample with the reference population or expected number of deaths predicted by a reference model, to obtain a more reliable information than the absolute value (Siegel et al., 2015).

The development of predictive or prognostic models is one of the focus in the literature, since it computes an estimated reference based on a predictor variable from the sample. Mortality prediction models have used severity scores as predictors. The SAPS (Simplified Acute Physiology Score), APACHE (Acute Physiology and Chronic Health Evaluation) and MPM_0 (Mortality Probability Model at zero hours) are the main severity scores used worldwide and have shown good prediction of mortality in several studies (Keegan & Soares, 2016).

Those severity scores have been updated throughout the years and currently SAPS and MPM_0 are in the third version (SAPS-3 and MPM_0-3), APACHE in fourth version (APACHE-IV) (Keegan & Soares, 2016). Information regarding the estimation of those scores is provided in Table 3. The choice of a severity score is important since it composes the references for the mortality indicator, and it may vary depending on where the sample is collected.

As shown in Table 3, APACHE and MPM have been applied to large samples of patients, however the SAPS-3 project has provided a standard equation and specific equations for estimating mortality in different world regions (Moreno et al., 2005). Furthermore, one important conclusion provided in Capuzzo et at. (2008) was that the prediction model using SAPS is limited only to the severity score and does not consider ICU characteristics.

Metnitz et al. (2009) stated that regional equations of SAPS-3 would be a good starting point for predicting mortality, however, in a country approach, a more specific score equation would be preferable. Prediction models with severity scores are evaluated in terms of Discrimination, usually the Area Under the Receive Operating Curve (AUC or aROC) and Calibration, with goodness-of-fit tests as the Hosmer-Lemeshow (H-L) (Keegan & Soares, 2016). Finazzi et al. (2011) assert that the H-L test does not provide information on the direction in which the observed-to-expected deviation is and can be influenced by the sample size of the cohort studied.

Table 3 - Comparison of Severity Scores Used as Mortality Predictors. Source: Afessa et al. (2007) and Keegan & Soares (2016)

Characteristics	SAPS-3 (Metnitz et al., 2005; Moreno et al., 2005)	APACHE-IV (Zimmerman et al., 2006)	MPM_0-3 (Higgins et al., 2007)
Study population	16,784	110,558	124,855
Study period	October 2002 – December 2002	January 2002 – December, 2003	October 2001 – March, 2004
Number of ICUs	303	104	135
Number of hospitals	281	45	98
Geographic regions	35 countries (5 continents)	USA	USA
Time of data collection	±1 h of ICU admission	24 h of ICU admission	Within 1 h of ICU admission
Variables in the model	20	142	16
Area Under Receiving Operating Curve	0.848	0.88	0.823
Hosmer-Lemeshow C statistic	14.29	16.9	11.62
Hosmer-Lemeshow H value	0.16	0.08	0.31
Standardized Mortality Ratio	1	0.997	1.018
Age	Yes	Yes	Yes
Length of hospital stay before ICU admission	Yes	Yes	No
ICU admission source	3	8	No
Type of ICU admission	Yes	Yes	Yes
Chronic comorbidities	6	7	3
Cardiopulmonary resuscitation before ICU admission	No	No	Yes
Resuscitation status	No	No	Yes
Surgical status at ICU admission	Yes	Yes	No
Anatomical site of surgery	5	No	No
Reasons for ICU admission/Acute diagnosis	10	116	5
Acute infection at ICU admission	Yes	No	No
Mechanical ventilation	Yes	Yes	Yes
Vasoactive drug therapy before ICU admission	Yes	No	No
Clinical physiologic variables	4	6	3
Laboratory physiologic variables	6	10	0

Therefore, the H-L statistics provide only an overall calibration measure, which can mislead the information on how the risk subgroups fit, evidencing a wrong miscalibration. In addition, the traditional calibration curves (expected risk *vs.* observed risk) using deciles also do not provide enough information regarding the calibration, since there is no statistical information about the fitness.

To address this problem, Finazzi et al. (2011) proposed the Calibration Belts, in which a confidence interval "belt" is estimated using the predicted risks. The curve of predicted risks is estimated by fitting a generalized polynomial logistic regression which relates the outcome with the logit transformed probabilities predicted by the severity score as predictor (Finazzi et al., 2011; Moralez et al., 2017). An example of a Calibration Belt is shown in Figure 1.

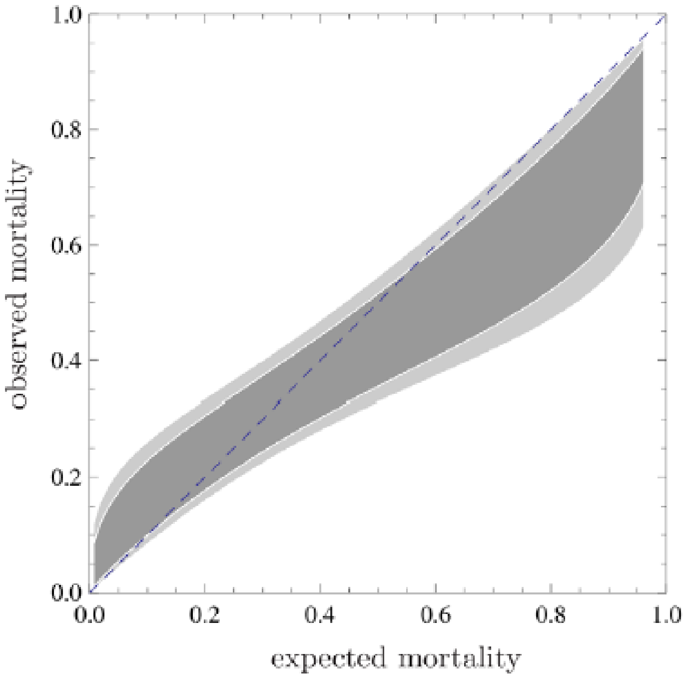

Figure 1 - Example of Calibration Belt. Source: Finazzi et al. (2011)

Those belts comprehend a confidence interval to estimate the degree of uncertainty of the calibration curve compared to the bisector line (predicted = observed), which results in a p-value from Wald's test. Furthermore, when analyzing the calibration belts, one can observe which region of the predicted curve contains, underestimate or overestimate the bisector line (Finazzi et al., 2011).

Poole et al. (2012) used the calibration belts to validate SAPS2 and SAPS3 curves in 103 Italian ICUs, which evidenced non-reliable patterns of both scores in predicting mortality risks; Moralez et al. (2017) computed the calibration belts to validate the SAPS3 and MPM_0 prediction curves in 48,816 Brazilian patients from the ORCHESTRA database, and evidenced that the SAPS-3 Standard Equation obtained a better calibration when predicting risks for Brazilian ICUs compared to the Central/South America Equation (Moralez et al., 2017; Silva Junior et al., 2010). More details on the calibration belts are described in Finazzi et al. (2011) and Poole et al. (2012).

It is important to notice that different distributions of SAPS-3 may result in better or worse fitness on the risk prediction equations. One limitation is that prediction risk models have presented overestimation of high severity cases and underestimation of low severity patients in several countries (Moreno et al., 2005), which impacts directly the observed-to-mortality indicators.

To achieve higher calibration, the literature sometimes perform a first-level customization of the prediction risks equations, in which the coefficients of the severity score, as re-estimated, using the logit transformed probabilities from the original equations as the predictor variable, and the patient’s outcome as the response; the second-level customization considers new predictor variables to re-estimate the response (Moreno et al., 2005). Although this may improve the fitness of the predicted risks, it does not solve all miscalibration problems (Moreno et al., 2005).

The main mortality indicator is the Standardized Mortality Ratio (SMR). For a certain group or ICU, it is calculated by dividing the ratio between observed number of deaths and the expected number of deaths, as expressed in (1) (Siegel et al., 2015).

$$SMR = \frac{Observed\ No.\ of\ Deaths\ or\ Mortality\ (O)}{Expected\ No.\ of\ Deaths\ or\ Mortality\ (E)} \tag{1}$$

The expected deaths can be obtained using a reference population adjusted by a risk factor - direct standardization -, or from a mortality prediction model – indirect standardization (Pouw et al., 2013). If SMR is greater than 1, it means that the ICU presents high mortality rate, since its number of deaths is greater than the expected, and if SMR is lower than 1 the ICU mortality is lower than the expected (Siegel et al., 2015).

For a single ICU, those thresholds may indicate a good or bad performance, however, as the SMR depends on an expected number of deaths, it may vary strongly with the prediction reliability of mortality risks (Metnitz et al., 2009). The has been the main indicator of mortality used in literature as it provides a good overview of a group or ICU’s performance regarding care outcome, and as it is easy to calculate (Metnitz et al., 2009). In the comparison among different ICUs, the SMR should be carefully assessed since it relates significantly with the hospital case-mix or severity scores (Pouw et al., 2013; Siegel et al., 2015).

2.2
Resource Use Indicators

Resource utilization has been presented as a secondary approach for evaluating ICU performance. One primary related outcome to resource use is cost. In the perspective of planning, costs would be a significant variable in decision making for the ICU or the Hospital administration, however, it is not a very accessible variable and its measurement may vary depending on the type of treatment and the case-mix presented (Rothen & Takala, 2008). The use of cost can be seen in Shwartz et al. (1995) and Lone et al. (2016).

When costs are not possible to obtain, studies have considered the ICU Length-Of-Stay (ICU-LOS) as a surrogate measure of resource utilization, adjusted by the mortality-risk of a severity score such as the SAPS-3 (Rothen et al., 2007; Rothen & Takala, 2008). The use of resources is directly associated with the efficiency of treatments, hence, the larger the length of a treatment (LOS), the higher is the resource utilization rate.

Although ICU-LOS is an easy variable to compute, its use can imply in high variability and possible bias, since it is directly influenced by the discharge criteria and patient transfer (Salluh et al., 2017). Thus, one must analyze it carefully before proceeding with the estimation of the expected resource utilization.

Rothen et al. (2007) used expected ICU-LOS of surviving patients to estimate the expected ICU-LOS, adjusted by risk, with the objective of computing the Standardized Resource Use (SRU), an indicator analogous to the SMR, expressed in equation (2).

$$SRU = \frac{Observed\ Resource\ Use\ (O)}{Expected\ Resource\ Use\ (E)} \tag{2}$$

The use of SRU is relevant as a standardized relation, between the observed and expected use of resources in an ICU. However, Rothen et al. (2007) stated that the ICU-LOS of patients who died must also be accounted, since they have required a high use of workforce in the treatment process.

To provide a more adequate expected value of ICU-LOS, other studies have provided regression models, using predictor variables from patients, such as the severity score and patient comorbidities, as in Niskanen et al. (2009), and Kramer & Zimmerman (2011). Other approach was the Weighted LOS, in which early days of hospitalization have higher weight compared to last ones, and the expected weighted hospital LOS is calculated with the severity score and percentage of surgical patients. (Hubert et al., 2007; Rapoport et al., 1994; Rothen & Takala, 2008)

Furthermore, the difference between observed and expected LOS is also used as a measure of relative resource use (Nathanson et al., 2007; Rothen & Takala, 2008). Similarly, scores for measuring the workload and use of staff, as the Therapeutical Intervention Scoring System (TISS) (Miranda et al., 1996) and the Nine Equivalents of Nursing Manpower (NEMS) (Miranda et al., 1997), have also been used to evaluate resource use rates, though those scores are not always accessible.

2.3
Benchmarking Methodologies

Different approaches to analyze indicators or to perform benchmarking have been widely used in cohort studies. The calculation of mortality indicators composes the primary objective of studies that evaluate different ICUs with the purpose of understanding possible variables that may influence the outcomes. Therefore, we consider three perspectives for analysis of performance: 1) SMR and SRU efficiency matrix that is traditionally used; 2) Rankability for ICUs, which comprehends the reliability of an indicator and obtaining performance groups; 3) Risk Profiles analysis, which provides a broader overview of the mortality risks per severity score to complement the mortality analysis.

2.3.1
Efficiency Matrix

The use of one indicator may result in biased analysis since the ICU is observed in just one perspective. Hence, the assessment of two or more indicators is recommended (Metnitz et al., 2009). Rothen et al. (2007) have provided an Efficiency Chart (or Efficiency Matrix), wherein the ICUs are plotted regarding SMR and SRU, as shown in Figure 2.

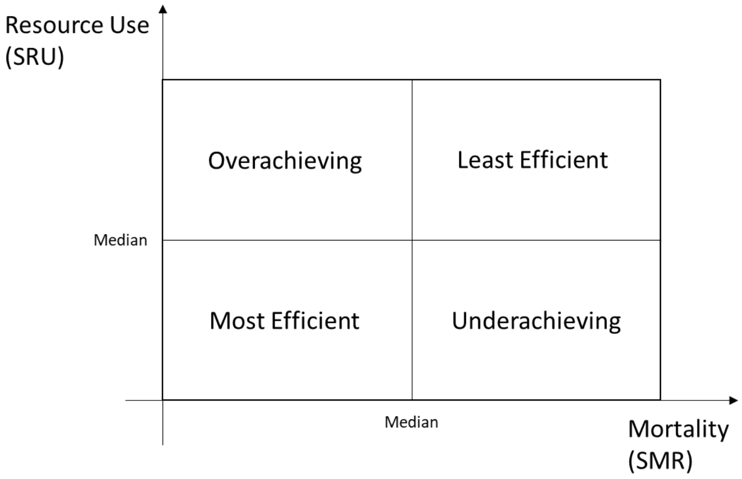

Figure 2 - Efficiency Chart/Matrix for Performance Analysis. Source: Adapted from Rothen & Takala (2008) and Salluh et al. (2017)

The reference lines relate to the overall medians of SMR and SRU from the sample, which provide a classification approach regarding the combination of its performance indicators. In Quadrant 1 (lower left), both indicators are below the mean, hence, it is considered as “most efficient” units. Conversely, in Quadrant 3 (upper right), there are the “least efficient” units and Quadrant 2 (upper left) comprises of the “overachieving” units, with high resource use and lower mortality. Quadrant 4 (lower left) comprehend the “underachieving” units, which present lower use of resources and high mortality (Rothen et al., 2007; Salluh et al., 2017)

Despite the limitations of both SMR and SRU, this chart is a good approach for evaluating a network of ICUs, especially if they are homogeneous or if they belong to the same organization. One can perform the analysis comparing the values of each ICU to the median SMR and median SRU of total ICU. Furthermore, Rothen & Takala (2008) claimed that the use several dimensions for evaluating performance , however, mortality outcome and resource use still provide a simple and fast overall view.

The efficiency chart can also be analyzed with different ICU characteristics. Rothen et al. (2007) verified clinical rounds and the presence of emergency departments in hospitals to be significant variables in most efficient ICUs. This approach was used in Soares et al. (2015), with the ORCHESTRA database, which concluded that there was significant variability in resource use, and most efficient ICUs in Brazil were characterized as private, with training programs, and using daily checklists.

2.3.2
League Tables and Rankability

League or Rank tables have been used to classify and rank ICUs (or hospital departments) in terms of a specific indicator. Regarding ICU benchmark, the rank tables provide information on which ICU performs best or worst in terms of mortality (SMR), along with the corresponding confidence interval.

Although Rank Tables are one straightforward approach to identify the benchmark within a set of ICUs, they may lack some reliability to infer if ICUs from close ranks are indeed different since their SMR confidence intervals overlap. Verburg et al. (2016) stated that rank tables ignore possible uncertainties due to sample size, which can imply that the rank or classification of an ICU may not be based on its real performance on patient care.

Therefore, van Dishoeck et al. (2011) proposed the Rankability, based on Goldstein & Spiegelhalter (1996), which corresponds to an indicator of how reliable a rank is. The Rankability created a signal-to-noise ratio, comprising two components: the heterogeneity and the uncertainty (van Dishoeck et al., 2011; Verburg et al., 2016). The heterogeneity comprehends the variance between ICUs (the signal), which corresponds to the true differences between ICUs (or hospitals) due to quality of care, management or processes. The uncertainty is the variance within ICUs (the noise) and represents the variability to design a certain rank to a hospital by chance.

Therefore, the Rankability indicator is calculated as expressed in equation (3) (van Dishoeck et al., 2011):

$$Rankability = \rho = \frac{\tau^2}{\tau^2 + \sigma^2} \tag{3}$$

Where ρ is the Rankability indicator that ranges from [0,1] and it can be expressed in percentage as well; τ^2 is the heterogeneity component and σ^2 is the uncertainty component.

The heterogeneity is computed as the variance from a random effects model, using the ICU as the random intercept, which accounts the differences in hospitalizations for each unit (Verburg et al., 2016). The uncertainty is calculated as the median standard error from the estimates of a fixed effects model, which considers the units as a categorical predictor of the outcome (van Dishoeck et al., 2011; Verburg et al., 2016).

van Dishoeck et al. (2011) calculated the Rankability from nine quality indicators, which ranges from 37% to 71%. Verburg et al. (2016) were the first to apply the Rankability to evaluate the Risk-Adjusted Mortality Rate (RAMR) in a set of ICUs. When the indicator is related to mortality, such as the SMR, the fixed and random effects model are logistics regressions, using ICUs as the random effect and the predictor variable, respectively, and the predicted risk from a model as an offset variable to adjust for severity (Verburg et al., 2016).

A high Rankability corresponds to a good classification of ICUs, however there has not been an ideal interval range for classifying a “reliable” ranking. Some authors indicate 75% as a minimum value for considering a good Rankability (Goldstein & Spiegelhalter, 1996). Verburg et al. (2016) adopted a minimum of 95% Rankability, which resembles the common significance level used in statistical hypothesis tests.

Verburg et al. (2016) showed that the Rankability increases as more patients are considered, being a longer period of evaluation (i.e. one year versus three years) or when grouping ICUs. The same authors then proposed a clustering approach for grouping ICUs considered “similar” and defined the number of clusters to be the one that maximizes the heterogeneity. Hence, a smaller group of similar units could be easily assessed instead of all population (individual ICUs), which is a good approach since it deals with the sample size problems of different ranks and allows the analysis of performance groups.

2.3.3
Risk Profiles

In the analysis of performance, the SMR stands out as the most widely indicator for mortality. However, as previously mentioned, the literature states limitations when using this indicator to compare ICUs mainly due to the behavior of the predicted risks from severity scores (Metnitz et al., 2005). In addition, Moreno et al. (2010) asserts that the SMR is a constant single-value to describe the performance from a case-mix of a unit, which limits the analysis since in fact an ICU may perform differently depending on the severity scores range.

Considering the assumption of performance variation depending on the severity score, Moreno et al. (2010) proposed the analysis of a Risk Profile of an ICU, which contains the predicted risk for each severity score on a unit. Therefore, one can evaluate how a certain ICU is performing on its low or high-risk groups compared to the reference.

The Risk Profile is computed as follows (Moreno et al., 2010): considering risk prediction model from a severity score (SAPS3, APACHE, etc.), firstly, a first level customization of this equation for each ICU is performed, to obtain the predicted risks from each specific unit, considering their case-mixes. Hence, predicted risks from the overall sample (all ICUs) are estimated.

Therefore, the predicted risk of each ICU is divided by the overall risks estimated on the whole sample from that respective severity. This results in a Risk Ratio indicator that comprehends on how well an ICU is performing for each predicted risk from the overall sample (Moreno et al., 2010).

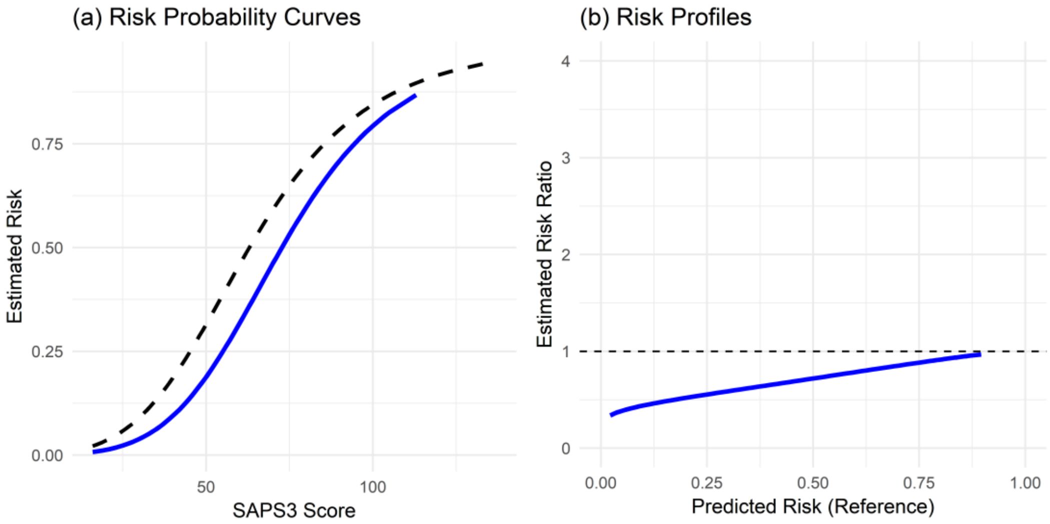

Figure 3 - Example of Risk Profile and Predicted Risks Curves of an ICU (blue line). Source: Based on (Moreno et al., 2010)

An example of the prediction and risk ratio curves is presented in Figure 3. In Figure 3 – (a) "S" curve from the logistic regression overall probabilities (black dashed line) and the curve estimated from the equation customized to a certain ICU (blue line). The ratio of the second by the first curve designs the curve in Figure 3 – (b). High risk ratios mean that the ICU has a higher mortality risk compared to the reference and may not be performing well for that range of severity group (Moreno et al., 2010).

In this sense, instead of a single value, there is all the span of mortality risks compared to the reference and it is possible to understand how each performance group deals with the mortality risks (Moreno et al., 2010).

2.4
Research gaps and questions

The evaluation the ICUs or Hospitals does not present a clear way of computing the outcomes (Rothen & Takala, 2008). Hence, the approaches to defining the correct benchmark try to compute valid expected values for comparing ICUs simplifying the overview of its performance, using indicators and the SMR and SRU. Keegan & Soares (2016) and Salluh et al. (2017) assert those indicators are still going to be used and updated with the purpose of providing improved accuracy in performance evaluation, especially with the use of big data tools. In addition, psychosocial and institutional outcomes have not presented a standard indicator, which could assist the planning analysis of the organization, though they were presented in intervention studies, as in Cavalcanti et al. (2016).

Brown et al. (2014) concluded that there has been no "best" measure among different indicators, being their performance directly related to the outcome they measured. Process measures have shown to be related to ICU aspects. Thus, they are not very sensitive to severity adjustment, while mortality measures are related to patients (Brown et al., 2014). Moreover, LOS measures have shown good performance and are related to costs. However, they are sensitive to bed availability (Brown et al., 2014).

Therefore, we observed that analysis of performance and benchmarking are still ongoing and important topics in healthcare systems, since they can provide a wide and deep understanding on how a unit is performing, their weaknesses and what are the best practices to achieve high quality of care. The limitations regarding performance analysis can imply in research opportunities to obtain better indicators or improve current ones. In addition, improved methods for benchmarking or analysis of performance can result in better control and monitoring of the process within a unit.

In this context, this research was based on the following research questions, which can still be considered for future works on benchmarking or analysis of performance on healthcare systems.

a) "What is analysis of performance in Intensive Care Units?"

This is the main research question that guides the entire study, as the literature present limitations and research gaps to develop new methods and discussions regarding the topics. Therefore, particularly in the ICU context, we considered this research question as the first step to understand the current state-of-art regarding performance indicators and benchmarking, which led to the specific research questions that guided the analysis in this study and its sample, as follows:

b) "Are the severity scores good predictors of Hospital Mortality?"

The analysis of performance should consider a good calibration of the severity scores predicted risks with the observed mortality in each ICU. In addition, the SMR limitations should be accounted and evaluated adequately to compare different units.

c) "How should resource use indicators be considered?"

As there is not a gold standard variable to compute the resource utilization parameters, different approaches should be considered and related to the mortality indicator. Therefore, the objective is to consider the SRU, estimated by the LOS, and relate them to the performance indicators and other outcome variables.

d) "What aspects of infrastructure and organization in ICUs are related to high/low performance?

Finally, aspects and characteristics of the ICU can correlate with a good or bad performance. Hence, as we provide information regarding the organization of each unit, the objective was to identify patterns among the aspects and the performance groups.

3
Research Methods

This research comprehended the performance analysis in a Brazilian ICU Network considering the indicators currently provided in the literature as well as the set of variables corresponding to the patient clinical status and outcomes, and the ICU characteristics.

In this sense, according to Vergara (2009), this study can be classified as an applied research regarding its purposes, since it comprehends the analysis of real data collected from ICUs to explore and provide solutions to problems; and also an explanatory research, since it aims to expose factors that may explain certain outcomes. Regarding its ends, this research can be considered as bibliographic, since it is based on a literature review regarding the main methods and applications related to the theme; as experimental, since there is the manipulation and control of certain variables to identify possible implications in the outcomes.

All the calculations were performed using R Software 3.4.4, with packages *caret* for regression analysis, *lme4* for random-effects modelling, *cluster* for clustering procedures, *dplyr* for database operations, *ggplot2* for plotting, and the *givitiR* to obtain the calibration belts, which was developed by the GiViTI (*Gruppo Italiano per la Valutazione degli interventi in Terapia Intensiva* - Italian Group for the Evaluation of the Interventions in Intensive Care Medicine).

For this purpose, the following steps were performed:

a) Data Collection and Preparation

In this step, raw data was provided from the Checklist Intervention Study (CHECKLIST-ICU) performed by Cavalcanti *et al*. (2016). We considered the information from the study population from two main databases: Patients and ICUs. A brief overview of the databases can be seen in Table 4.

Table 4 - Dataset Overview

Dataset	Variables/Categories
Patient	Admission (Age, Genre, Comorbidities, SAPS-3 score, SOFA score, Admission Date, Reason for Admission) 15-day Treatment / Intervention study (Patient treatment monitoring, infections, use of catheters, use of mechanical ventilation) Outcomes – ICU & Hospital (Vital Status, Date of Outcome, Length of Stay - LOS)
ICU	Characteristics, Organization, Infrastructure, Human Resources, Assistive Resources, Material Resources, Assessment, Work Procedures, Patient Transportation, Risk Management, Infection Prevention & Control

The first database had data regarding the admission status of each patient, the treatment information from a period of 15 days, and the outcomes of patients in the 60th day of hospitalization. The second database comprehends the characteristics of each ICU and its respective hospital as well as their resources (human and assistive), safety procedures, treatment equipment and organization. The latter was obtained from a questionnaire composed of multiple choice questions regarding the adherence of ICUs to the guidelines proposed by the Ministry of Health, known as the RDC-7 Standard, in (MINISTÉRIO DA SAÚDE, 2010). A brief overview on the categories, number of items (questions) and the alias used to refer to each variable is provided in Table 5 and the sub-items are in Appendix I.

Table 5 – Categories from ICU database

Category	Alias	Number of Questions
Organization	ORG	6
Physical Infrastructure	PF	3
Human Resources	HR	9
Assistive Resources	AR	17
Work Procedures	WP	2
Patient Transportation	PT	2
Risk Management	RM	2
Infection Prevention and Control	INF	3
Evaluation	EV	10
Material Resources	MR	10

Patient database comprised a total of 13,635 records, and the ICU database presented information on 130 distinct ICUs. To ensure the quality of data, the patient's database was first adjusted to follow the evaluation standards in Cavalcanti et al. (2016) in terms of eligibility:

- Adult patients: age must be greater or equal than 18;
- Only patients included in the study after 48h of ICU admission date should be considered;
- Patients with high probability of death between 48h and 72h after ICU admission date should not be included;
- Patients with hypothesis of or confirmed brain death, should not be included;

Furthermore, it is important to state that this study analyzed the patient’s outcomes and ICU performance within a total hospital length of stay of 60 days (ICU admission date + length of stay) as in Cavalcanti et al. (2016). Hence, patients who were discharged or continued in hospital in the 60th day were considered as “surviving”.

Thus, considering the final eligible patients, the Patient and ICU databases were combined. In Figure 4, one can observe the database preparation procedure for obtaining the final data. Finally, the three databases displayed complete information from a total of 12,100 patients, within 116 ICU.

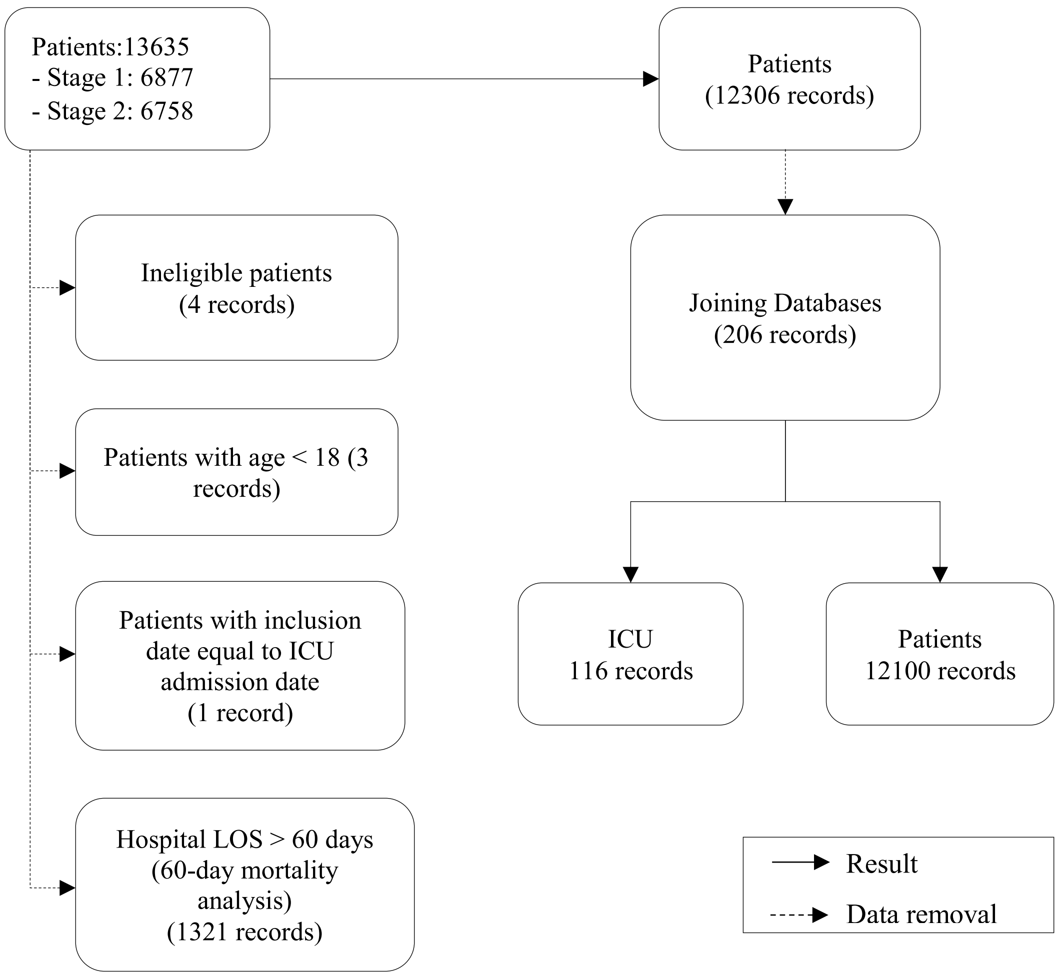

Figure 4 - Dataset Preparation Procedure

b) Descriptive Analysis

For obtaining an overall view on the databases, descriptive analysis was used to display statistics regarding the sample of patients and ICUs. In this step, the characteristics of patients and ICUs were computed. Information was displayed in tables, providing the quantity and percentage related to total number of cases for categorical variables, median and interquartile range (1st quartile – 3rd quartile) or mean and standard deviations (SD) for quantitative variables.

Charts were also used to display relation between each characteristic and the mortality outcome. Thus, bar plots were considered for analyzing the mortality level for each category of patient and ICU/Hospital. The quantitative variables regarding severity scores and bed quantities were also categorized according to the common statistics in literature.

c) Prediction of Mortality Risks and Calibration

For the mortality indicator, the expected number of deaths was predicted using the SAPS-3 equations provided in Moreno et al. (2005), both Global/Standard Equation (SAPS3-SE) and Central/South America Equation (SAPS3-CSA), since this latter corresponds to the world region containing Brazil. The corresponding equations are expressed in equations (4) and (5). The estimated probability of death is obtained by equation (6).

$$Logit(SE) = -32.6659 + \ln(SAPS\ 3 + 20.5958) \cdot 7.3068 \qquad (4)$$

$$Logit(CSA) = -64.5990 + \ln(SAPS\ 3 + 71.0599) \cdot 13.2322 \qquad (5)$$

$$P(Death) = \frac{e^{Logit}}{(1 + e^{Logit})} \qquad (6)$$

To evaluate the fitness of the prediction curve, we computed the calibration belts, proposed by Finazzi et al. (2011), for both SAPS3-SE and SAPS3-CSA original. The evaluation considered the best calibration by analyzing the behavior of the predicted curve and its confidence interval in terms of containing the bisector line.

To improve the calibration of both curves, we performed a first-level customization on the SAPS-3 equations. This process comprehends the re-estimation of the predictor coefficient (SAPS-3) by fitting a new logistic regression considering the outcomes of the sample (survival, death) and the logit transformed probabilities estimated from the original equations. Therefore, we selected the curve with best calibration regarding its adequacy to the bisector line as the main reference for the remaining analysis.

d) Analysis of Performance

For this step, performance indicators were applied to the patient and ICU databases. In this sense, mortality and resource use outcomes were considered as two-dimensional assessment of each ICU, using the hospital outcome and the ICU LOS for each patient. Hence, we calculated the SMR and SRU for each ICU and designed the performance matrix from Rothen et al. (2007) and Salluh et al. (2017).

The SRU was calculated considering the expected value of LOS obtained from surviving patients, adjusted by decile of the severity score, as proposed in Rothen et al. (2007), and expressed in equation (7).

$$E[Total\ Resource\ Use] = \sum_{i=\ Deciles} Average\ LOS\ (surviving)_i \cdot No.Patients_i \qquad (7)$$

With the information of SMR and SRU, we plotted the efficiency chart (SRU x SMR) for evaluating the performance of the ICU network and observed the most efficient and least efficient ICUs from the databases (Salluh et al., 2017).

In addition, we grouped the ICUs considering the characteristic variables provided in the ICU database to observe the overall distribution of the performance groups for each aspect. To analyze the significance of each characteristic, we tested the normality of the SMR and SRU distribution with the Shapiro-Wilk Test. Then, we tested for differences in the mean in the categories of each aspect variable considering the One-Way ANOVA and the t-Test, when the indicators are normally distributed, or the Kruskal-Wallis and Mann-Whitney Tests, when the indicators are non-normal, for variables with more than two categories and two categories, respectively.

Moreover, we related the ten institutional categories with the performance indicator by computing their correlation. Each category was computed as a score, which represented the total number of total and partial positive adherence of a certain ICU to the practices in that category.

We performed a univariate linear regression for each category considering the SMR and SRU as response variables to obtain individual statistical significances of each category. Hence, we performed a multivariate analysis with the same response variables: firstly, we used the LASSO regression to select the most important variables and we computed a multivariate linear regression with the chosen variables.

The overall statistical significance level considered was 0.05 for the regression analysis and statistical tests.

e) Rankability and Risk Profiles

To complement the evidences from the efficiency matrix, we calculated the Rankability for the overall ICUs using the method provided by Verburg et al. (2016) considering the reliability of the SMR. Hence, we performed the clustering procedure for the maximum heterogeneity to evaluate performance groups on the ICUs. For each clustering technique, we calculated the Rankability, uncertainty and heterogeneity values.

We considered four main common clustering procedures to assess possible differences between the clusters as there is no standard technique for this procedure: Hierarchical Clustering (HC) – Agglomerative and Divisive, K-Means and K-Medoids. For the Agglomerative HCs, we analyzed four linkage functions: Ward's Distance, and Complete, Single, Average Linkage, representing the "Agglomerative Nesting" (AGNES), the Divisive HC corresponded to the Divisive Analysis Clustering ("DIANA"), and for the K-Medoids, we used the Partition Around Medoids (PAM) algorithm. A brief description on the clustering techniques is available in Appendix II and one can find more details in Khanmohammadi et al. (2017) and Reynolds et al. (2006).

For the best clustering results regarding heterogeneity, we computed the median SMR and SAPS-3 as well as the SRU to evaluate how the groups are composed, mainly in terms of SMR, SRU, median SAPS-3 (case-mix) and the number of ICUs.

Hence, with the same best results we computed the risk profile curves of each cluster to evaluate how they are performing on their case-mix's predicted risks. We used the method proposed by Moreno et al. (2010) using the clusters as groups instead of each individual ICU.

f) Reporting Evidences

Finally, the main evidences from the analysis of performance in the sample are discussed regarding their importance and possible causes. We identified the main correlations patterns between the variables and the outcomes (response) as well as the evidences regarding the best or worst performance groups, and the clustering procedure.

We recall that this analysis was performed in a database used for the clinic trial analysis in (CAVALCANTI et al., 2016). Hence, this study was limited on the variables and data collected from the reference work, for both ICUs and Patients.

4
Descriptive Analysis

We present a descriptive analysis on the two databases used for this study. Firstly, we report the main aspects regarding the patients and their outcomes, with the variables informed in their dataset. Then, we present information on the characteristics of the ICU analyzed in this study. For both databases, we also analyze the lethality patterns with bar plots for each aspect.

4.1
Database of Patients

We present a first overview of the Patient's database comprising the characteristics of this study's population. Overall descriptive information comprising the status of the patient at admission were obtained as well as the statistics separated per hospital outcome. One can observe the descriptive statistics in Table 6.

In this sample, the overall hospital lethality was 35%. Considering the total of patients, we observed that most admissions were Medical (70.5%), and "Postoperative care" was the main reason for admission (21.6%), followed by "Respiratory Failure (Except Sepsis)" (15.5%) and "Sepsis" (12.5%). The category "Others" aggregates reasons with smaller frequency of patients. Furthermore, most patients had "Cancer treatment" as comorbidities (8.1%).

Regarding the outcomes categories, we verified that the patients with "Death" outcome were on average older (64 years old) than patients with "Discharge" (57 years old), which confirms the common assumption of age being an important factor for mortality risk. Moreover, patients who died stayed longer on average in ICU (12 days) than discharged patients (9.3 days) and presented a higher median SAPS-3 (61) and SOFA (6) scores than discharged patients (SAPS-3 = 45, SOFA = 2). Hence, we may evidence that patients who died were indeed those who required more treatment and used more resources.

Table 6 - Descriptive Statistics - Patients Dataset

Variables	Total (n = 12100)	Outcome (Hospital) Discharge (n= 7870, 65%)	Death (n=4230, 35%)
Age, mean (SD)	59.6 (19)	57 (19.1)	64 (18.1)
Female sex, No. (%)	5661 (46.8)	3717 (47.2)	1944 (46)
ICU LOS, mean (SD)	9 (8.4)	7.5 (6.9)	12 (10)
Post-ICU LOS, mean (SD)	7.1 (9.5)	9.3 (9.5)	-
Type of Admission, No (%)			
Medical	8529 (70.5)	5164 (65.6)	3365 (79.6)
Elective Surgery	2035 (16.8)	1686 (21.4)	349 (8.3)
Emergency Surgery	1536 (12.7)	1020 (13)	516 (12.2)
Reason for ICU Admission, No. (%)			
Postoperative Care	2616 (21.6)	2092 (26.6)	524 (12.4)
Respiratory Failure (Except Sepsis)	1873 (15.5)	966 (12.3)	907 (21.4)
After cardiorespiratory arrest	167 (1.4)	74 (0.9)	93 (2.2)
Neurological	1480 (12.2)	952 (12.1)	528 (12.5)
Hepatic	181 (1.5)	73 (0.9)	108 (2.6)
Gastrointestinal	331 (2.7)	225 (2.9)	106 (2.5)
Sepsis	1518 (12.5)	692 (8.8)	826 (19.5)
Shock (Except Sepsis)	131 (1.1)	64 (0.8)	67 (1.6)
Cardiovascular	1420 (11.7)	1156 (14.7)	264 (6.2)
Renal/metabolic	509 (4.2)	299 (3.8)	210 (5)
Hematological	92 (0.8)	51 (0.6)	41 (1)
Others	1782 (14.7)	1226 (15.6)	556 (13.1)
Comorbidities, No. (%)			
Cancer treatment	978 (8.1)	530 (6.7)	448 (10.6)
Heart	842 (7)	509 (6.5)	333 (7.9)
Cirrhosis	289 (2.4)	126 (1.6)	163 (3.9)
AIDS	372 (3.1)	179 (2.3)	193 (4.6)
SAPS3, median (IQR)	50 (39 - 63)	45 (36 - 55)	61 (51 - 74)
SOFA, median (IQR)	3 (1 - 7)	2 (0 - 5)	6 (3 - 9)

The "Discharge" outcome category was mostly composed of Medical admissions (65.6%), and the Reason for admission with most patients was "Postoperative care" (26.6%), followed by "Cardiovascular" (14.7%) and "Respiratory failure" (12.3%), similar to the overall evidences. The "Death" outcome mostly comprised of Medical admissions as well (79.6%), however it had a higher frequency of Emergency Surgery admissions (12.2%), compared to the surviving patients; and the main Reason for admission was "Respiratory Failure" (21.4%), followed by "Sepsis" (19.5%), and "Neurological" (12.5%). Therefore, we can notice that patients who died were admitted with diseases commonly known for providing a high severity health status such as Sepsis.

We considered the SAPS-3 as the main severity score for the analysis due to its broader approach regarding the patient compared to the SOFA, which is mainly used to score the organ failure process, and its use in performance analysis for predicting mortality risk. The distribution of the SAPS-3 scores in the sample is shown in Figure 5, for the total number of cases and deaths.

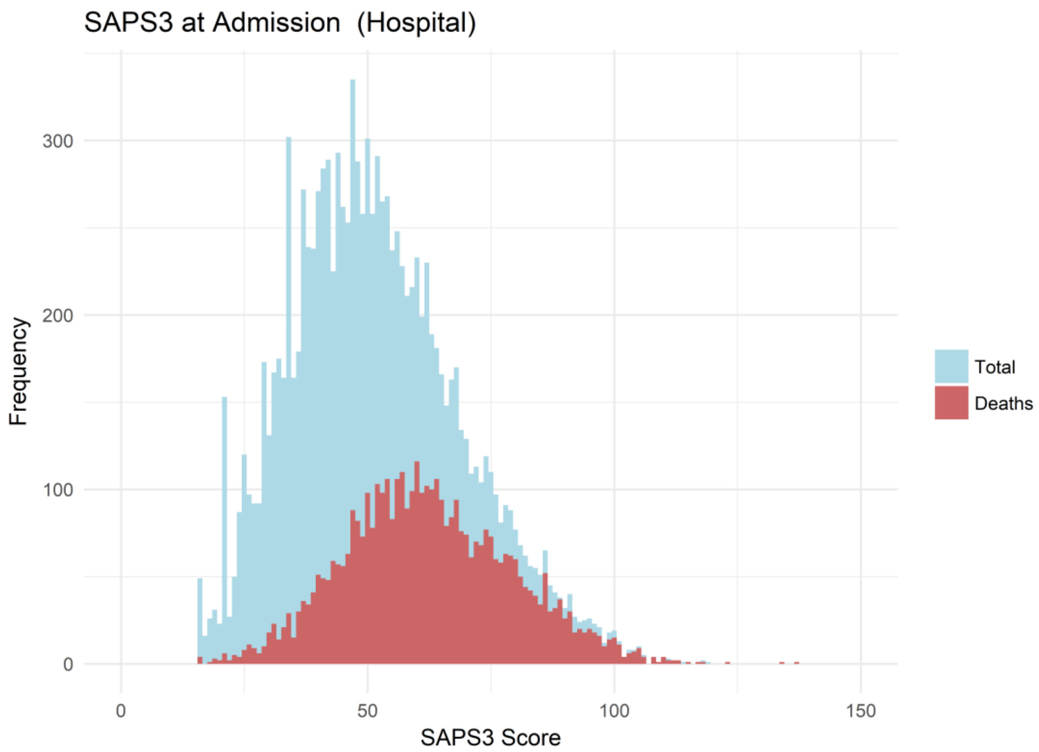

Figure 5 - SAPS-3 Score Distribution

We observed that SAPS-3 distribution resembles a lognormal distribution, with extreme values in the right tail. Patients who died tend to present higher severity scores, as previously verified in Table 6. In this sample, the maximum SAPS-3 score was 137 and the minimum was 16, which is the overall minimum value in the SAPS-3 scale.

Furthermore, we computed the frequency and the lethality plot for each Reason for admission. The data is shown in Figure 6. In this plot, we sorted the categories in ascending order of lethality.

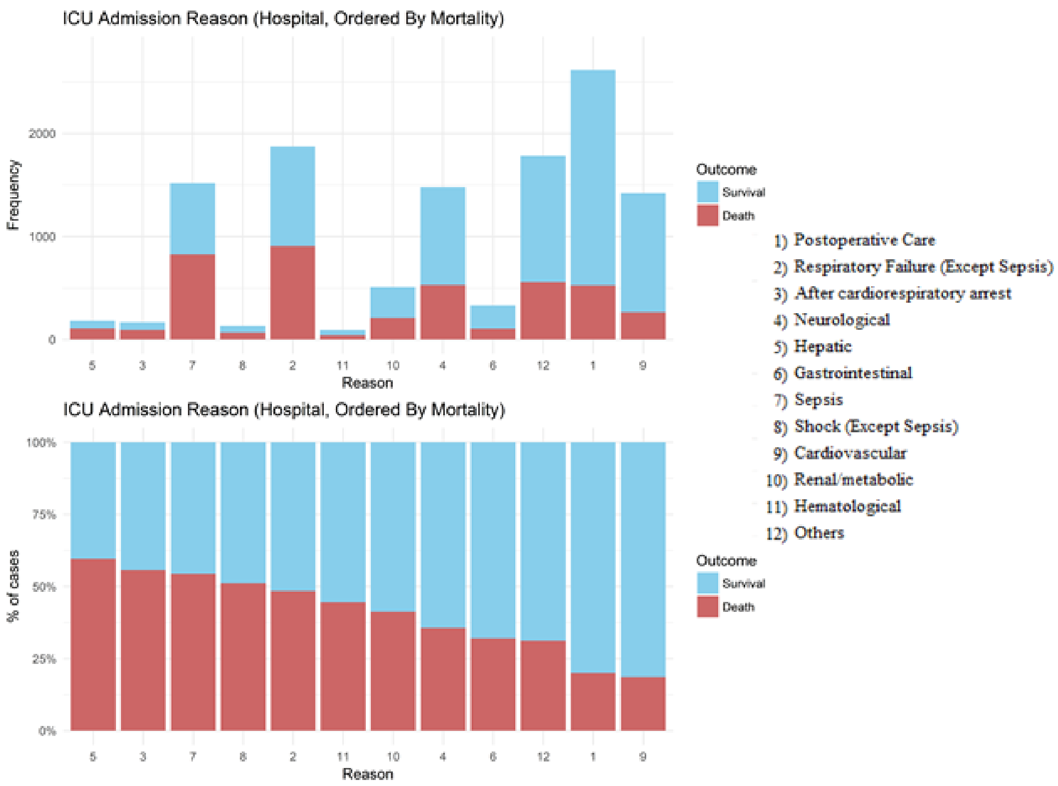

Figure 6 – Reasons for admission outcome frequency and lethality plots

One can observe that “Postoperative care” presented one of the lowest lethality rates, in spite of being the category with most of the patients. “Hepatic” was the Reason for admission with highest lethality rate, followed by “After cardiorespiratory arrest”, however represented by few cases. “Sepsis” presented a considerable lethality rate and frequency of patients, which was observed mainly in patients who died as well as the “Respiratory Failure”.

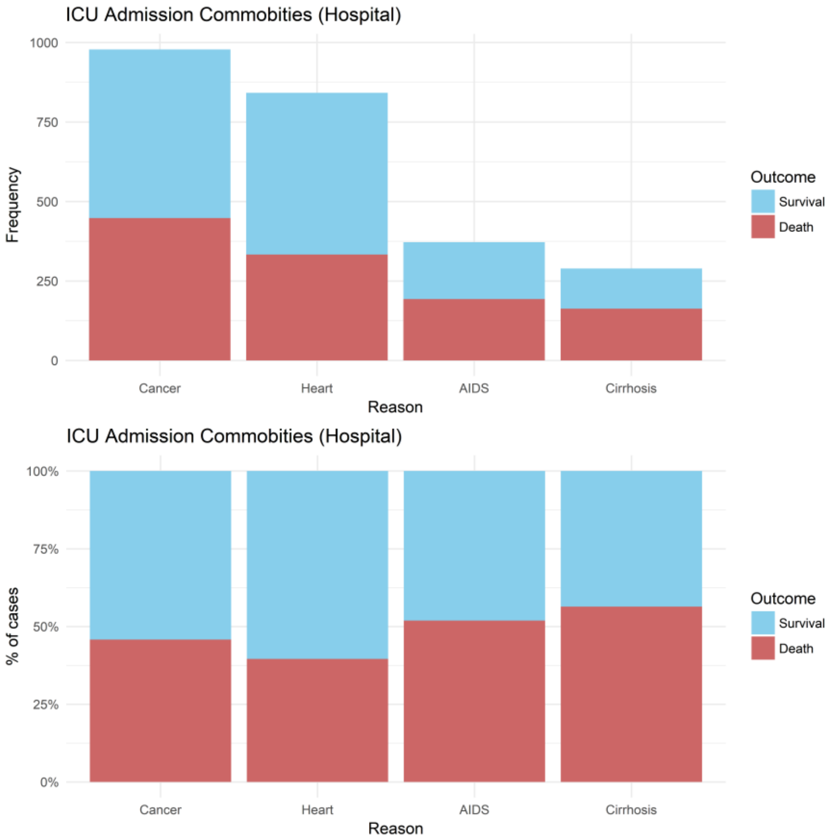

Figure 7 – Comorbidities outcome frequency and lethality plots

We also observed the outcomes behavior per comorbidities as shown in Figure 7. Cirrhosis was the main comorbidity in terms of lethality, which may be related to the "Hepatic" reason for admission in Figure 6, however it presented the lowest frequency. "Heart" was the comorbidity with the lowest lethality, which may follow the lowest lethality for the "Cardiovascular" reason for admission as well.

4.2
Database of ICUs

Furthermore, descriptive information regarding the ICUs related to the patient dataset was computed as observed in Table 7. The information comprehended the quantity of ICUs belonging to a certain characteristic provided in the dataset.

One can verify most ICUs belong to public administration (47%), to hospitals in which their beds are mostly dedicated to the Brazilian Unified Health System (*Sistema Único de Saúde* – SUS, 75%), to Tertiary hospitals (78%) and to General Hospitals (86%). Furthermore, most of ICUs are Mixed (78%) and only 33% are in Academic Hospitals. In terms of Hospital bed quantity, most of ICUs belong to hospitals with considerable number of beds (101 -500, 71.55%), and with 11 – 30 ICU Beds (68.1%), which also comprehends the categories with highest lethality.

Table 7 - Descriptive Statistics - ICU Dataset

ICU Characteristics	n = 116, %	Patients	Lethality (%)
Complexity, No. (%)			
Primary	2 (1.72)	242	24.38
Secondary	23 (19.83)	2453	34.65
Tertiary	91 (78.45)	9405	35.31
Hospital Type, No (%)			
General	100 (86.21)	10500	34.31
Specialized	16 (13.79)	1600	39.19
Bed-Assignment, No. (%)			
SUS	87 (75)	8981	38.48
Private Healthcare	29 (25)	3119	24.82
Hospital Admin., No. (%)			
Public	55 (47.41)	5522	40.42
Private nonprofit	32 (27.59)	3475	35.68
Private for-profit	29 (25)	3103	24.43
Academic Hospital, No (%)	38 (32.76)	3891	37.32
ICU Type			
Mixed	90 (77.59)	9395	34.92
Medical	14 (12.07)	1424	32.58
Surgical	5 (4.31)	532	27.82
Specialized	7 (6.03)	749	44.99
No. of Beds - Hospital, Median (IQR)			
< 50	3 (2.59)	292	21.23
51 - 100	15 (12.93)	1559	29.57
101 - 500	83 (71.55)	8746	36.28
500 +	15 (12.93)	1503	35.53
No. of Beds - ICU, Median (IQR)			
< 10	20 (17.24)	2076	34.87
11 - 30	79 (68.1)	8325	35.77
31 - 50	12 (10.34)	1174	31.35
50 +	5 (4.31)	525	30.48

The lethality information per Hospital Complexity is shown in Figure 8 – (a). We observe that even though ICUs from Tertiary Hospitals contain most of the patients, the lethality per category is quite similar. Moreover, ICUs from Tertiary, and even Secondary hospitals may have a wide range of resources compared to Primary hospitals, which may also justify the large number of patients.

Regarding Hospital Types, General Hospitals contained of most patients and higher lethality rate, as observed in Figure 8 – (b). However, the lethality from both categories were slightly different, with Specialized hospitals being a little higher than General.

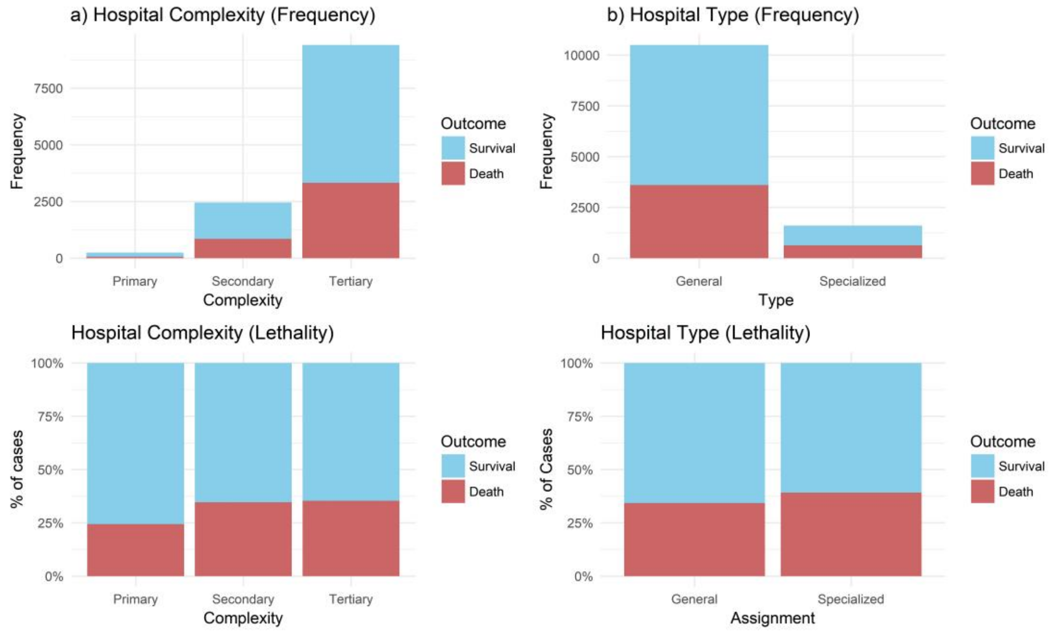

Figure 8 – Hospital Mortality: (a) Hospital Complexity; (b) Hospital Type

The lethality information regarding the Bed-Assignment categories and the Hospital Administration are shown in Figure 9.

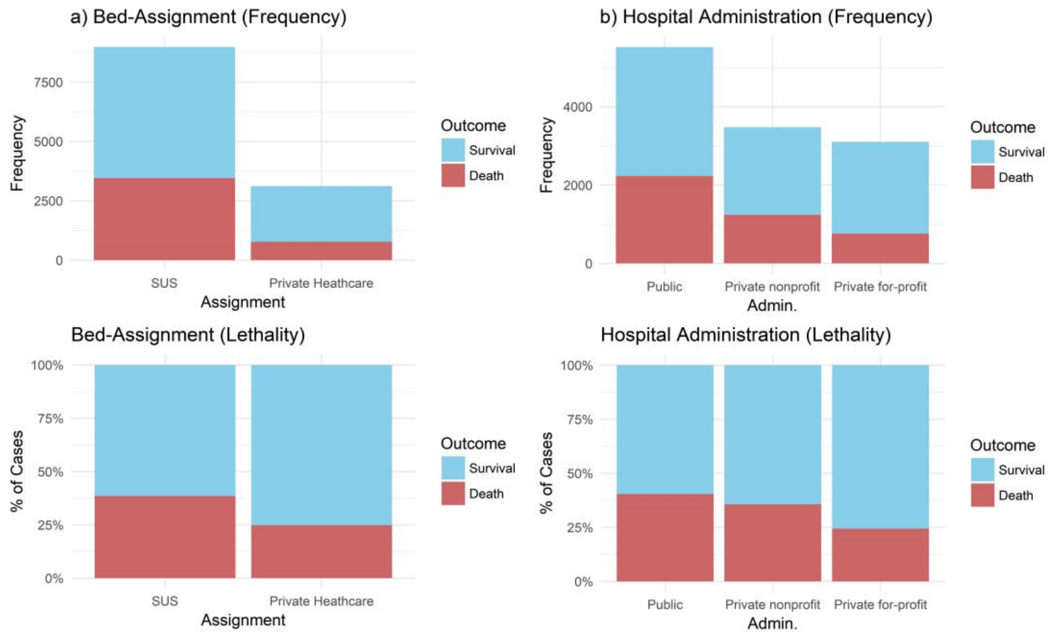

Figure 9 – Hospital Mortality: (a) Major Bed-Assignment; (b) per Hospital Administration

On the first variable, "SUS" contained most patients and higher lethality compared to Private Healthcare, as shown in Figure 9 – (a). This pattern is also assigned to the Hospital Administration variable, as observed in Figure 9 – (b), which reports that ICUs from Public hospitals also present high mortality percentage, compared to for-profit hospitals. In Brazil, the "SUS" belongs to a public administration, hence there is a possible high correlation between those variables, and it has more limited resources available, compared to private healthcare, which can influence in a worse outcome.

Finally, we reported the information on the lethality plot for ICU Types in Figure 10.

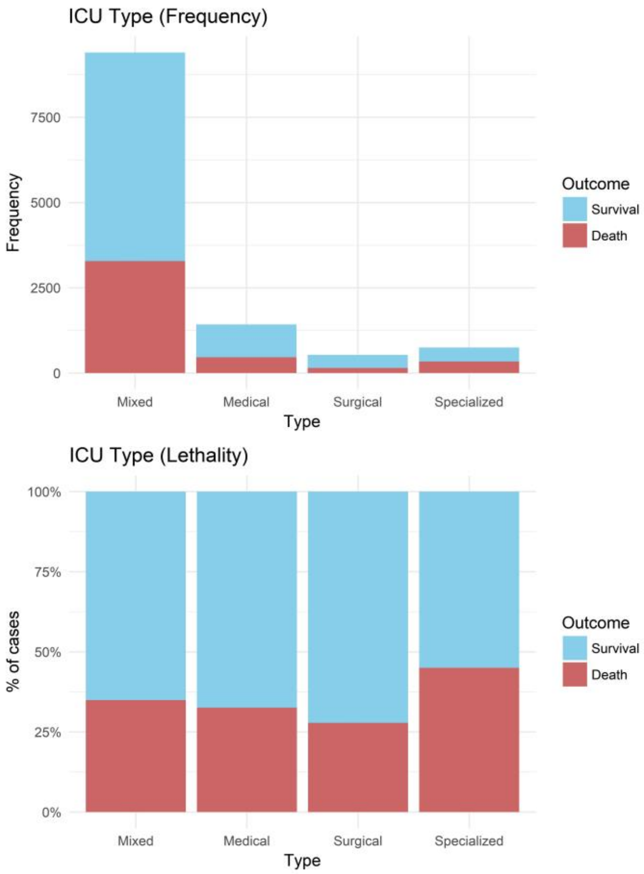

Figure 10 – Hospital Mortality - ICU Type

Although the Mixed ICUs contained most of the patients in the study sample, Specialized ICUs reported the highest lethality rate, with a few patients. Surgical ICUs have the lowest lethality rate and number of patients.

Therefore, in the patient's database the severity score SAPS-3 is the primary data to be used in the performance analysis, which, by definition, comprises information on age, comorbidities, reason for admission and possible tests performed in the patients. In terms of ICU, the hospital administration and bed-assignment evidence some correlation with the outcome (lethality).

5
Analysis of Performance

We conducted the analysis of performance in four sequential steps which are presented in this section. First, we evaluated the calibration of the SAPS-3 equations for predicting mortality risk provided in the literature to identify the reliability of the prediction, its implication in the SMR and to choose an adequate prediction model for the study. Then, we calculated the efficiency matrix with SMR and SRU and after the relation between the efficiency to the main aspects from ICU provided in the database. Next, we calculated the Rankability for the SMR and we performed a cluster analysis using this reliability indicator to evaluate the different ICU groups in different clustering techniques. Finally, we provided the Risk Profile for each ICU cluster to identify their performance in terms of mortality risks.

5.1
Calibration of SAPS-3 Equations

We calculated the predicted mortality risks using the original SAPS-3 equations for the Global and the Central/South America references: SAPS3-SE and SAPS3-CSA. Using the calibration belts approach, we obtained the calibration curves plots in Figure 11.

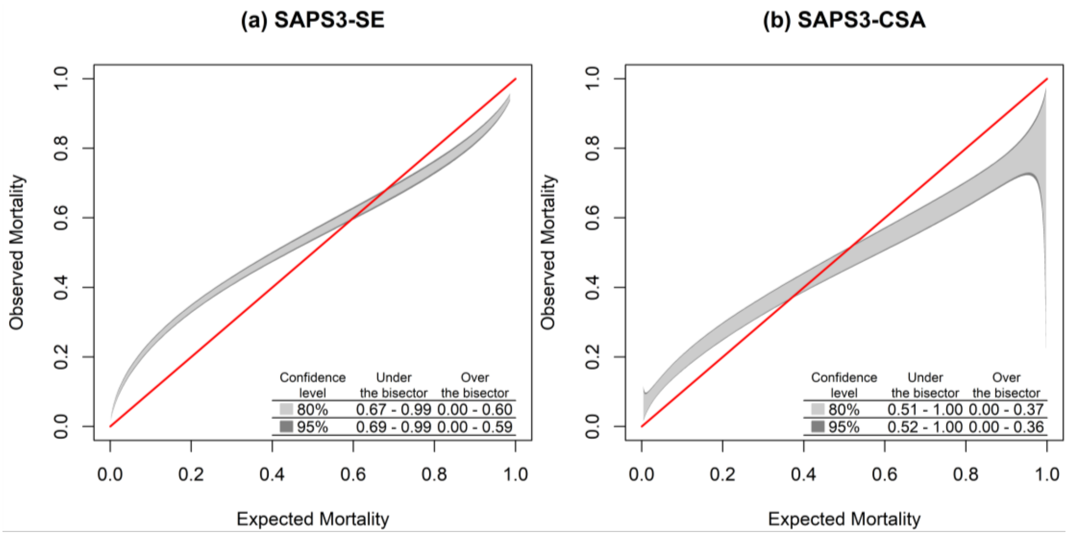

Figure 11 – Calibration Belts for Original SAPS-3 Equations: (a) Standard Global; (b) Central/South America

One can observe that both equations appear to be quite miscalibrated with the lower predicted risks under the bisector curve (underestimation of low-risk patients) and the higher predicted curves over the bisector curve (overestimation of high-risk patients), which has been previously indicated in the literature.

Hence, we performed the first-level customization in both equations. The new curves are shown in Figure 12.

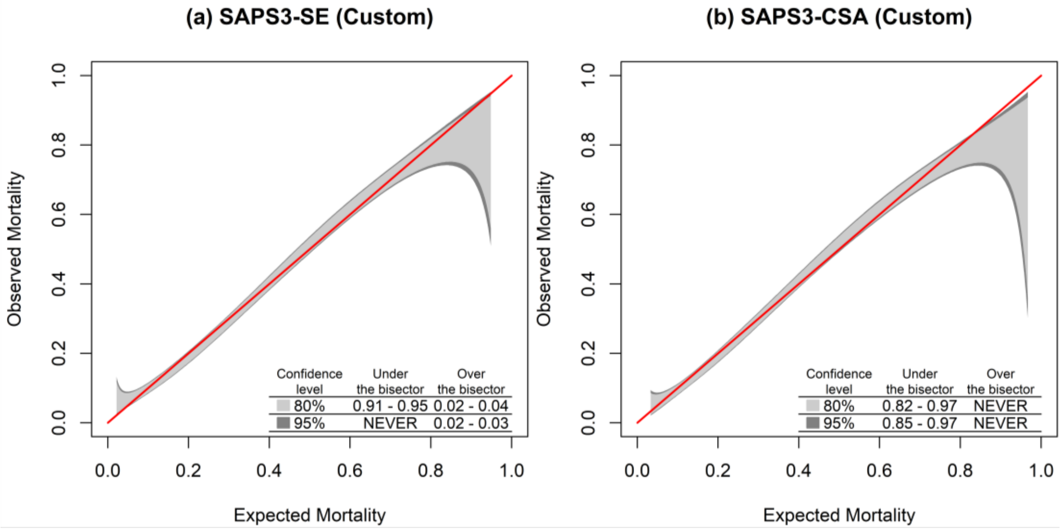

Figure 12 – Calibration Belts for Customized SAPS-3 Equations: (a) Standard Global; (b) Central/South America

The calibration provided a better fitness on the prediction curves. It is possible to observe that the bisector curve is mostly contained by the calibration belt in both curves, except for predicted risks close to zero. Hence, we selected the SAPS3-SE customized equation (SAPS3-SE/Custom) for this sample to compute the predicted risks and the next analysis since it can provide a global reference for the ICUs compared to the SAPS3-CSA.

The behavior of the calibration belts for the original equations imply that those equations must be carefully used in benchmark studies since there was a bias evidence. This comportment occurs since the SAPS3 equations were estimated from a different sample, with a slightly lower severity distribution compared to the sample from this study and with a ten-year difference. If one does not consider a proper calibration, the SMR and future analysis may result in wrong inferences and evidence on the studied sample.

5.2 Efficiency Matrix

With the customized SAPS-3-SE equation, we obtained the predicted risks for each score and we calculated SMR as well as the SRU indicator with eq. (7) for each ICU. Then, the Efficiency Matrix is shown in Figure 13, with the overall medians as the reference lines, as in Table 8.

Table 8 - Overall Reference values of SRU and SMR

Overall Reference	Median	IQR
SRU (Exp. Survival LOS)	0.99	0.72-1.25
SMR (SAPS3-SE/Custom.)	0.95	0.96-1.33

We observe that the median SMR and SRU are closer to the common reference 1.00 (expected = observed), which may represent a good balance of those indicators. This pattern also is resembled in the SRU IQR, while the SMR presents a slight shift towards higher values.

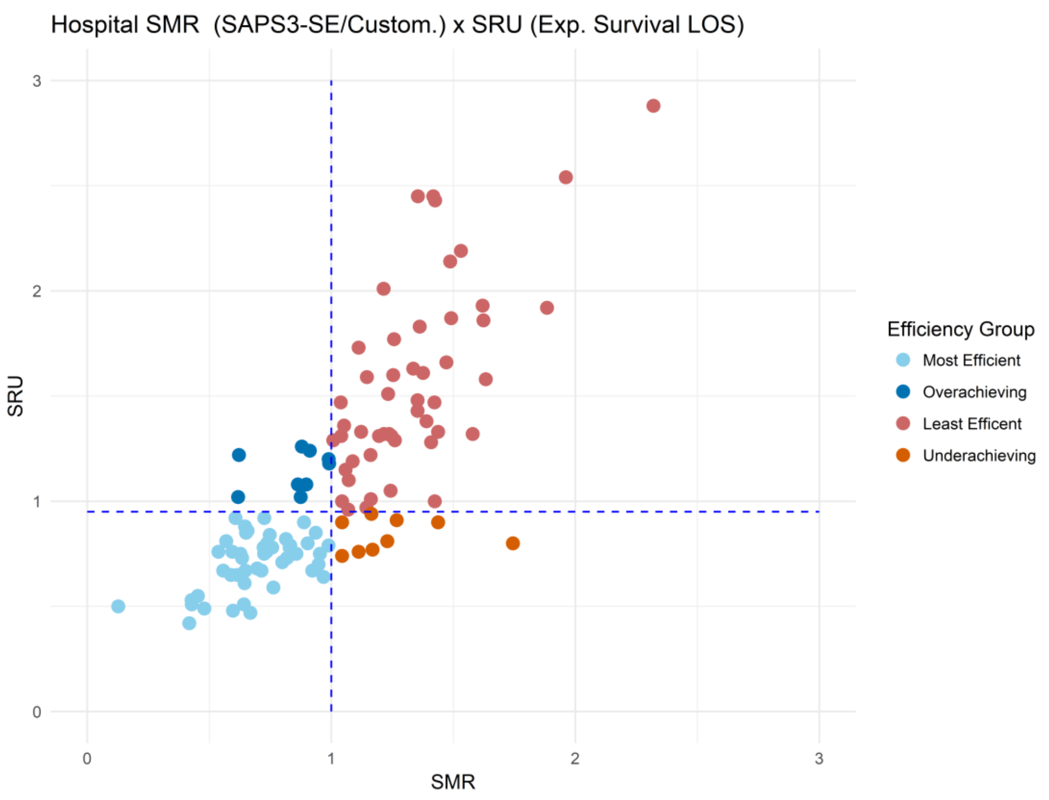

Figure 13 – Efficiency Matrix for the CHECKLIST-ICU sample

In the efficiency matrix, we can observe both indicators appear to correlate positively since the dispersion resembles a linear relationship. The “Least Efficient” category presents a wider dispersion of the ICUs compared to the “Most Efficient” group. We recall that both indicators have scales comprehending the interval [0,∞), which may justify the small dispersion of ICU points in the most efficient quadrant, which can only hold values in a smaller interval, while the least efficient quadrant does not have an upper bound.

Table 9 - Statistics of Performance Groups

Efficiency Groups						
Group	n	SRU (median)	IQR	SMR (median)	IQR	SAPS-3 Median
Most Efficient	49	0.75	0.65-0.79	0.72	0.78-1.03	51
Overachieving	9	1.18	1.08-1.22	0.88	1.10-1.29	48
Least Efficient	49	1.47	1.29-1.83	1.26	1.58-2.01	51
Underachieving	9	0.81	0.77-0.90	1.17	1.53-2.39	44

More information on the performance groups is shown in Table 9. The ICU with the best performance, prioritizing the mortality indicator, has SMR = 0.13, and SRU = 0.5, and the one with worst performance presents SMR = 2.32, SRU = 2.88. A smaller quantity of ICUs is in the "Overachieving" or "Underachieving" quadrants, while the remaining units are distributed in the main performance groups. One can also observe that the SAPS-3 in the most and least efficient group was similar, which indicates that for the same case-mixes we have ICUs performing differently due to possible organizational processes.

5.2.1
ICU characteristics and Efficiency Matrix

To observe how the different characteristics of the ICUs are distributed related to the efficiency group, we obtained the more relevant aspects by evaluating statistical differences among the categories in each variable. Firstly, we tested the normality of SMR and SRU individual distributions. Results are shown in Table 10.

Table 10 – Normality tests for the SMR and SRU distributions

Variable	Shapiro Wilk's Test	
	Estimate	P-value
SMR	0.98	0.05
SRU	0.90	**< 0.001**

The results from the Shapiro-Wilk's test rejected the null hypothesis for the SRU. Although the same should be considered for the SMR, the p-value was considerably low and, thus, we considered the non-normality as well. Therefore, as both indicators did not present strong evidence of normality, we computed the

Kruskal-Wallis and the Mann-Whitney tests for differences in the mean (for SMR and SRU) among than two categories, and between two categories, respectively. The results are presented in Table 11.

Table 11 – Results from the Tests on mean differences per ICU characteristic

Characteristics	Test	SMR			SRU		
		Estimate	DF	P-Value	Estimate	DF	P-Value
Complexity	Kruskal-Wallis	0.90	2	0.64	2.12	2	0.35
Hospital Administration		13.36	2	< **0.001**	12.00	2	< **0.001**
ICU Type		5.83	3	0.12	3.84	3	0.28
Hospital Bed Quantity		6.89	3	0.08	5.56	3	0.14
ICU Bed Quantity		6.10	3	0.11	4.18	3	0.24
Hospital Type	Mann-Whitney	792.00	-	0.95	607.50	-	0.12
Bed-Assignment		1769.00	-	< **0.001**	1736.00	-	< **0.001**
Academic Hospital		1419.00	-	0.71	1376.00	-	0.53

We verified that “Hospital Administration” and “Bed-Assignment” presented statistically significant differences among their categories (p-value < 0.05) regarding both SMR and SRU. This evidence may correlate with the lethality results observed in Figure 10. We can observe that the results from SMR and SRU were similar. “Hospital Bed Quantity” and “ICU Bed Quantity” did not rejected the null hypothesis, however they presented lower p-values, 0.08 and 0.11, respectively for the SMR, and “Hospital Type” computed the lowest p-value from the non-significant variables for the SRU.

Hence, we grouped the units in the efficiency matrix by assigning the categories from each variable related to a characteristic. We discuss mainly the significant variables and the grouping of remaining aspects are shown in Appendix III.

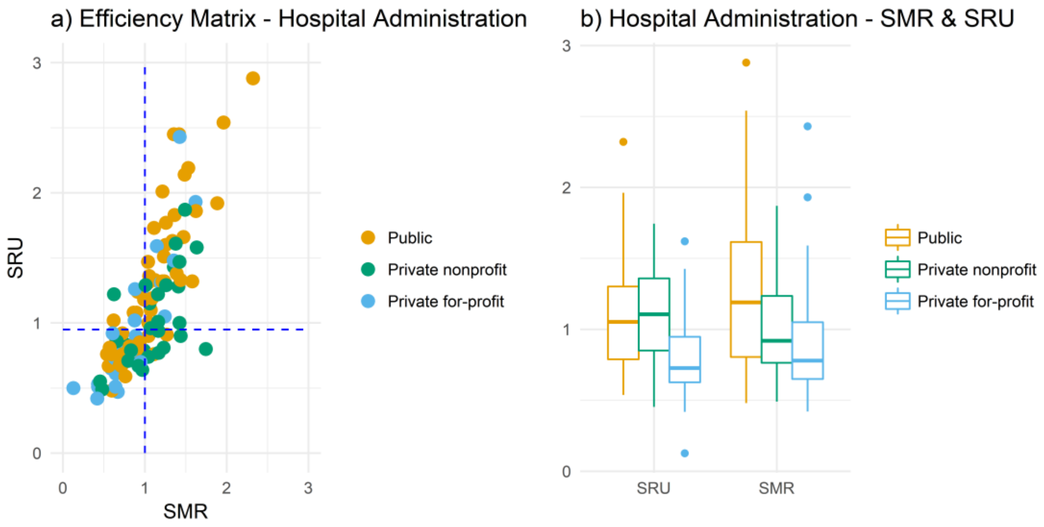

Figure 14 – Hospital Administration: (a) Efficiency Matrix; (b) SMR and SRU Boxplot

One can observe in Figure 14 – (a) that most ICUs from the least efficient quadrant belong to Public Hospitals. ICUs from Private For-profit hospitals are present towards the most efficient quadrant. Furthermore, the Underachieving quadrant comprises mostly of units from Private non-profit hospitals. This behavior is confirmed in Figure 14 – (b), we verify that the SMR values tend to be lower from Public to Private For-profit hospitals, while the SRU is higher in Private Non-profit. Therefore, we have evidence that the administration between Public and Private For-profit categories are indeed different and the former tends to have higher mortality and considerable higher use of resources compared to the latter.

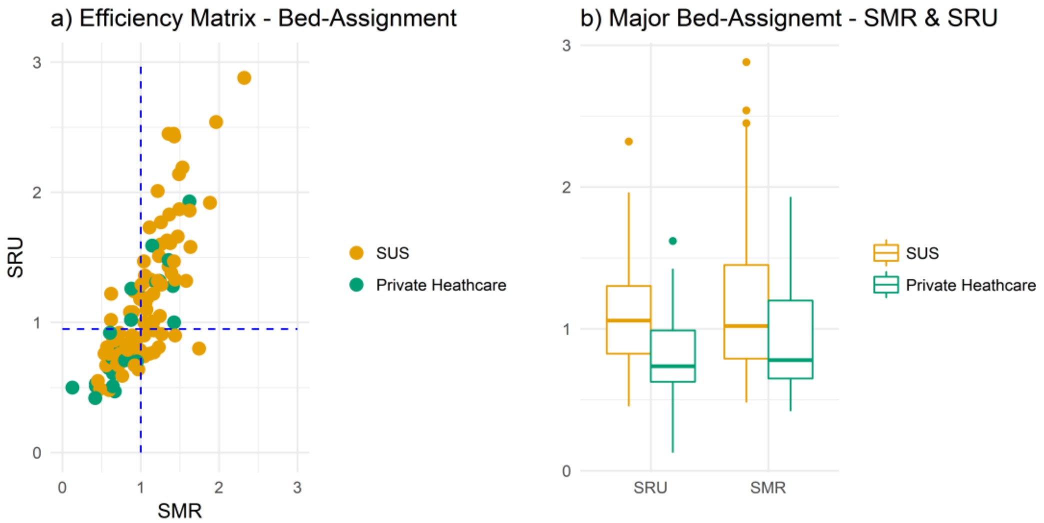

Figure 15 – Bed-Assignment: (a) Efficiency Matrix; (b) SMR and SRU Boxplot

The “Bed-Assignment” variable presents a similar pattern to the “Hospital Administration”, as shown in Figure 15 – (a). ICUs that assign most of their beds to the SUS compose most of ICUs from the least efficient group. ICUs related to Private Healthcare are concentrated mostly in the most efficient quadrant. In Brazil, ICUs or hospitals that attend more SUS patients than private healthcare generally belongs to public administration or private non-profit. Therefore, we also observe in Figure 15 – (b) that “SUS” category tend to have a higher resource utilization, and mortality; we also recall it comprises of 75% of the total ICUs in this study.

5.2.2
ICU institutional variables and Performance Indicators

We expanded the analysis to verify the relation between the performance indicators and institutional variables presented in the study’s sample. Considering the categories from the ICU dataset provided in Table 4 and the scores generated for each variable, we computed a correlation matrix as a first overview of the relationships among them, which is shown in Table 12. As we observed a high correlation between categories PI and INF, we combined both variables into “Infrastructure and Infection Prevention & Control” category, defined as “II”.

Table 12 – Correlation Matrix: SMR/SRU and Institutional Variables

	SMR	SRU	ORG	PI	HR	AR	WP	PT	RM	INF	EV	MR	II
SMR	1	0.80	-0.20	-0.23	0.04	-0.10	-0.09	0.06	-0.19	-0.26	-0.23	-0.02	-0.26
SRU	0.80	1	-0.11	-0.34	-0.07	-0.13	-0.10	0.06	-0.25	-0.37	-0.26	-0.17	-0.37
ORG	-0.20	-0.11	1	0.28	0.20	-0.06	0.09	0.26	0.31	0.22	0.28	0.23	0.27
PI	-0.23	-0.34	0.28	1	0.17	0.26	0.01	0.23	0.29	0.78	0.33	0.40	0.96
HR	0.04	-0.07	0.20	0.17	1	0.01	0.10	0.11	0.11	0.17	0.13	0.28	0.18
AR	-0.10	-0.13	-0.06	0.26	0.01	1	-0.11	0.04	0.08	0.25	0.01	0.13	0.27
WP	-0.09	-0.10	0.09	0.01	0.10	-0.11	1	-0.07	-0.03	0.07	0.05	-0.09	0.03
PT	0.06	0.06	0.26	0.23	0.11	0.04	-0.07	1	0.27	0.21	0.19	0.23	0.23
RM	-0.19	-0.25	0.31	0.29	0.11	0.08	-0.03	0.27	1	0.22	0.43	0.10	0.28
INF	-0.26	-0.37	0.22	0.78	0.17	0.25	0.07	0.21	0.22	1	0.32	0.37	0.92
EV	-0.23	-0.26	0.28	0.33	0.13	0.01	0.05	0.19	0.43	0.32	1	0.26	0.35
MR	-0.02	-0.17	0.23	0.40	0.28	0.13	-0.09	0.23	0.10	0.37	0.26	1	0.41
II	-0.26	-0.37	0.27	0.96	0.18	0.27	0.03	0.23	0.28	0.92	0.35	0.41	1

From this table, we also generated the correlogram in Figure 16 to visually identify the correlation patterns.

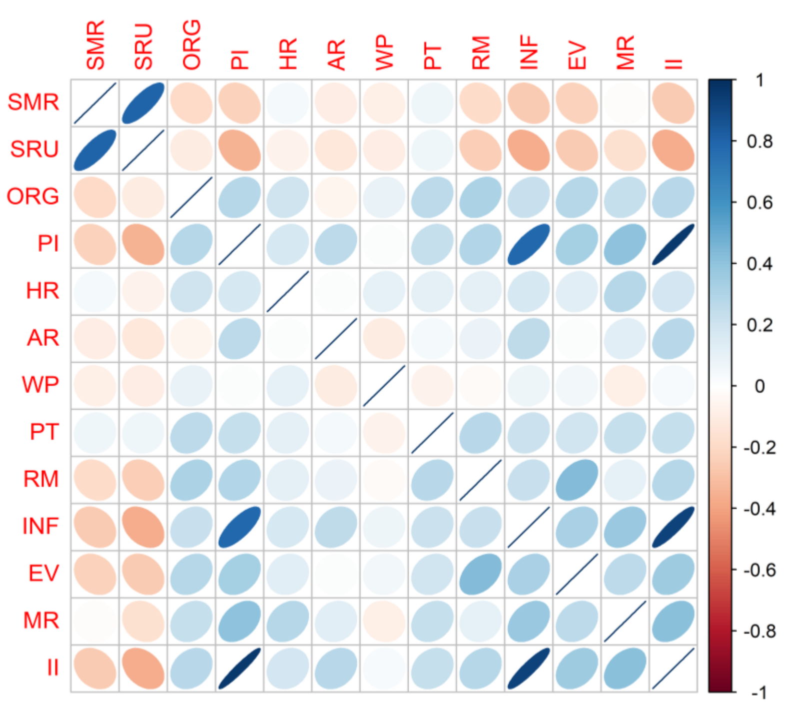

Figure 16 – Correlogram of Institutional variables and performance indicators

Table 13 – Results from the Univariate Regression

Categories/ Predictors	Response = SMR		Response = SRU	
	Estimate	P-Value	Estimate	P-Value
Organization	-0.03	**0.04**	-0.03	0.22
Physical Infrastructure	-0.10	**0.01**	-0.19	< **0.001**
Human Resources	0.02	0.70	-0.04	0.47
Assistive Resources	-0.02	0.27	-0.03	0.16
Work Procedures	-0.31	0.36	-0.49	0.28
Patient Transportation	0.04	0.50	0.05	0.55
Risk Management	-0.08	**0.04**	-0.14	**0.01**
Infection Prevention & Control	-0.15	**0.01**	-0.29	< **0.001**
Evaluation	-0.03	**0.01**	-0.05	< **0.001**
Material Resources	-0.01	0.82	-0.06	0.08
Infrastructure & Infection	-0.07	**0.01**	-0.13	< **0.001**

In the univariate regression, we observed that except from the “Organization” variable, all the significant categories for the SMR were the same for the SRU. Moreover, they were also the pairs with strong correlation. “Human Resources” and “Patient Transportation” have positive estimates, however they were not significant, and we noticed a small positive correlation coefficient in Figure 16.

We verify that “Infection Prevention & Control” presented the estimate with highest value, which was not evidenced when merged with the “Infrastructure” category. “Risk Management” presented the second highest value for the estimate.

Therefore, we evidence that presenting a good risk management routine, evaluation procedures as well as the presence of adequate infrastructure and infection control policies, individually, represented factors that influence in good performance of the studied ICUs.

To identify and select possible significant variables for the multivariate analysis, we performed a LASSO Regression. The results from this procedure identify variables that have non-zero estimates that composed a multivariate regression. Results are presented in Table 14.

Table 14 – Results from the LASSO and Multivariate Regression

Categories/ Predictors	LASSO Regression		Multivariate Regression			
	SMR	SRU	SMR		SRU	
	Estimate	Estimate	Estimate	P-Value	Estimate	P-Value
Evaluation	-0.004	0.00	-0.02	0.11	-	-
Infrastructure & Infection	-0.020	-0.05	-0.05	**0.04**	-0.13	**< 0.001**

The LASSO Regression model resulted on estimates for the “Evaluation” and the “Infrastructure and Infection Prevention & Control” for the SMR, and only the latter for the SRU, which represented the model with lower variance estimated by the procedure. When we estimated a multivariate regression model on the respective variables, we noticed that “Infrastructure & Infection" was indeed a significant institutional category for both SMR and SRU – for this indicator only this variable was considered from the LASSO regression model and it indicates that the more the ICU adheres to the its practices the better is the performance (negative estimate).

We recall that each category considers a score of the questions answered to compute the adherence to the best practice. Hence, each sub-category was assumed to have equal weights and each variable comprise of a different number of sub-categories. Therefore, for a ICU, the manager may consider different weights (priorities) for each item, as well as additional data to improve the assessment of performance.

5.3 Rankability and Clustering

Contemplating the analysis on the SMR, the overall Rankability indicator considering 116 ICUs was 80%, which is lower than the 95% threshold used in the literature. To improve the indicator, we grouped ICUs using different clustering techniques and considered the best number of clusters as the one that maximizes the heterogeneity. The plots regarding the evaluation of heterogeneity and uncertainty for each clustering procedure are provided in Appendix IV. As the SMR is strongly correlated to the SRU, we assume those results can also be considered for the performance groups.

Therefore, the information regarding the Rankability composition regarding the optimal number of clusters for each technique are presented in Table 15.

Table 15 – Rankability components for each clustering technique

Method		Heterogeneity	Uncertainty	Rankability	# Clusters
AGNES	Ward	1.159	0.008	0.993	11
	Complete	1.635	0.004	0.998	5
	Average	2.155	0.047	0.979	3
	Single	2.290	0.040	0.983	4
	DIANA	1.537	0.010	0.994	6
	K-Means	1.055	0.010	0.991	12
	K-Medoids	0.920	0.011	0.988	15

One can notice that the number of clusters varies considerably depending on the procedure we are using. All the clustering techniques achieved a high Rankability (Over 95%). The AGNES with Single linkage technique obtained the highest heterogeneity component, and the AGNES with Complete linkage presented the lowest uncertainty. The former merges cluster with lowest dissimilarity distance, hence, ICUs with closer SMR values tend to form a cluster, while the latter consider longer dissimilarity distances, which may include clusters that are considerably different.

K-Means and K-Medoids which are techniques that have a random approach in their procedures provided more clusters. There is not an optimal number of clusters to be evaluated performance groups. This difference in the total clusters among techniques evidence that there was not a stability on the clusters and one should look down into the characteristics of each cluster to evaluate the best approach. Therefore, we assessed information regarding the SMR, the number of ICUs, SRU, the Lethality and the SAPS3 Median values per cluster in each technique. We observe the SMR distribution per cluster in Figure 17.

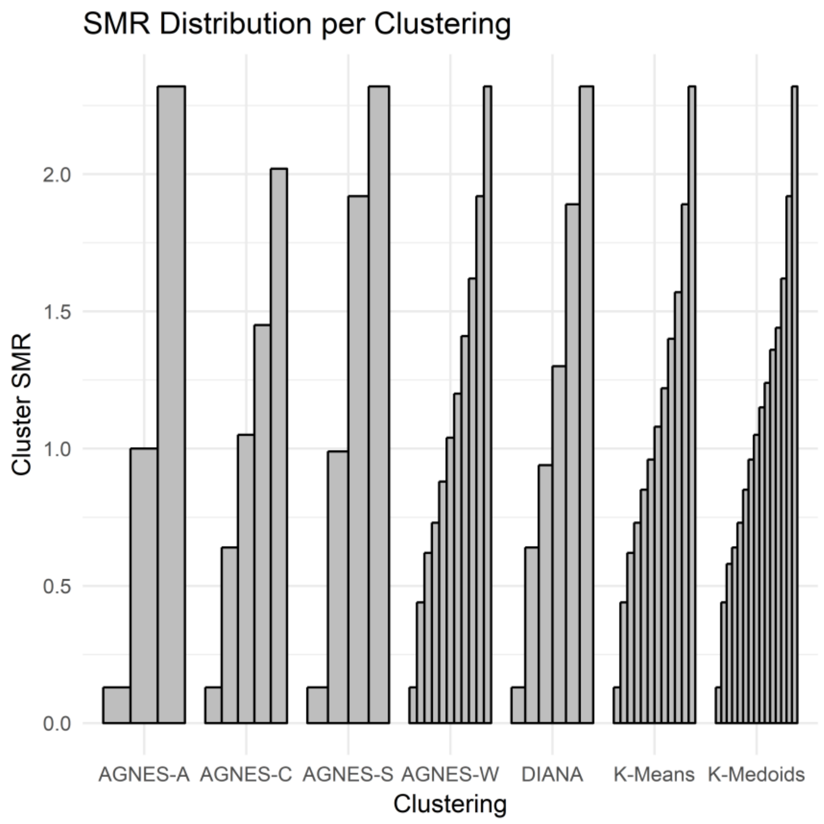

Figure 17 – Clustering procedure comparison: SMR per Cluster

The clusters within each clustering technique were proposedly ordered in terms of SMR. Hence the first clusters are the ones with lower SMR and the last clusters presented highest SMR. We observe that the difference between the first and last clusters is large, which implies in less overlapping of the SMR confidence interval, which promotes the higher reliability of the SMR ranks.

In addition, their difference is also large to the intermediate clusters, which would be expected as the clustering was performed considering only the SMR as the main variable. Hence, very high or low values of SMR that are extreme would be apart.

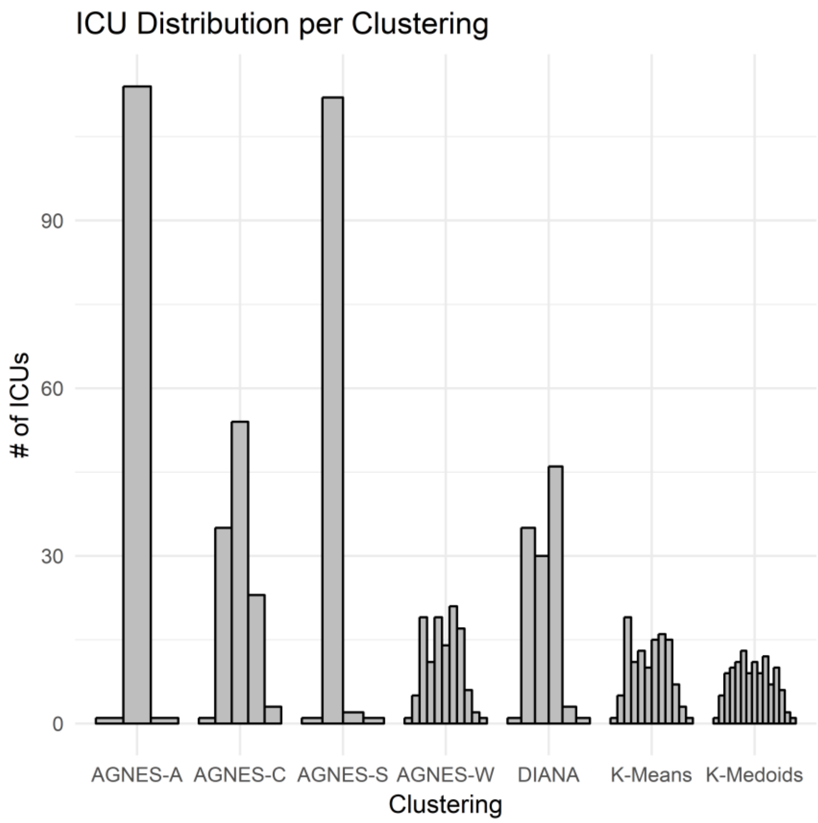

Figure 18 – Clustering procedure comparison: Number of ICUs per Cluster

In Figure 18, it is represented the comparison among the different clustering methods in relation the total number of ICUs that compose each cluster. It is possible to notice that the extreme clusters (first and last) are composed by a small number of units, while the intermediate clusters vary in number according to the technique. We evidenced there is not a balance in terms of number of ICUs for each clustering procedure.

K-Medoids appear to balance the number of ICUs per cluster, however 12 clusters could be a large evaluate performance groups considering only the SMR. Conversely, AGNES with Average Linkage (AGNES-A) presented three clusters, being two composed of one ICU each, and the remaining units composes a single cluster, which may not assist the evaluation of different performance groups due to the large difference in sample sizes. Hence, the clustering procedure has shown to be very sensitive to SMR values.

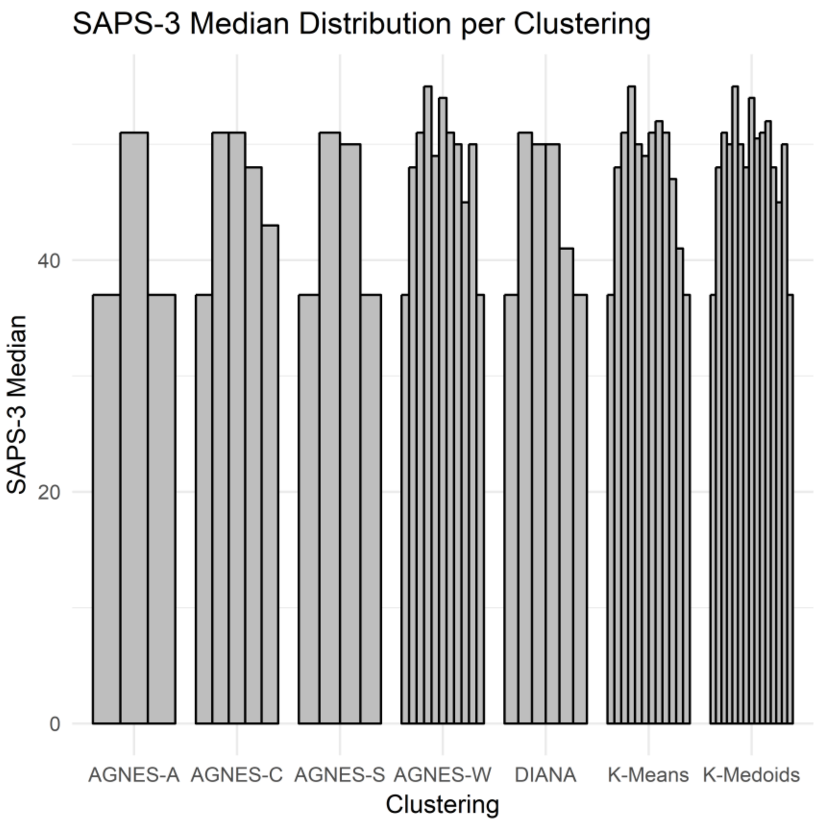

Figure 19 – Clustering procedure comparison: SAPS-3 Median per Cluster

In Figure 19, we computed the SAPS3-Median to verify the case-mix within each cluster. We observed that the two clusters (extreme clusters) comprised of composed with ICUs with the same SAPS-3 Median values (case-mix), which comprehended patients with lower mortality risks compared to other clusters. Therefore, we evidence in this example, the high SMR did not follow exactly the case-mixes of high severity patients, as one should expect since an elevated SAPS-3 score has a high mortality risk.

Therefore, we identified that clusters of ICUs considering the Rankability can vary strongly with the SMR distribution in the sample. Furthermore, the severity scores did not necessarily follow the SMR values and should be considered in the evaluation as well. We evidenced that the best and worse clusters in terms of SMR (and SRU) were composed by the same number of ICUs and similar case-mixes, which may suggest that the difference in performance is also explained in the treatment of its case-mix, especially if it is composed by low-severity patients. More information on clusters per technique is reported in Appendix V.

5.4
Risk Profiles Evaluation

To evaluate the case-mix mortality span, we designed the risk profiles for each clustering technique as well as the probability curve estimated by the SAPS-3 equations. Following the considerations in the Rankability, the clusters' ordination is maintained.

We have evidence that the case-mix also varies in each cluster for all procedures, and clusters with high SMR were not composed only of high-severity cases. In addition, the extreme clusters presented the same case-mix in almost all clustering techniques. Therefore, we used the Risk Profiles to evaluate the risks within each clustering techniques. For the analysis, we discuss on the clusters provided by the AGNES-A and the K-Medoids techniques, which comprised of the lowest and highest number of clusters, respectively. The plots for the remaining clustering techniques are provided in Appendix VI.

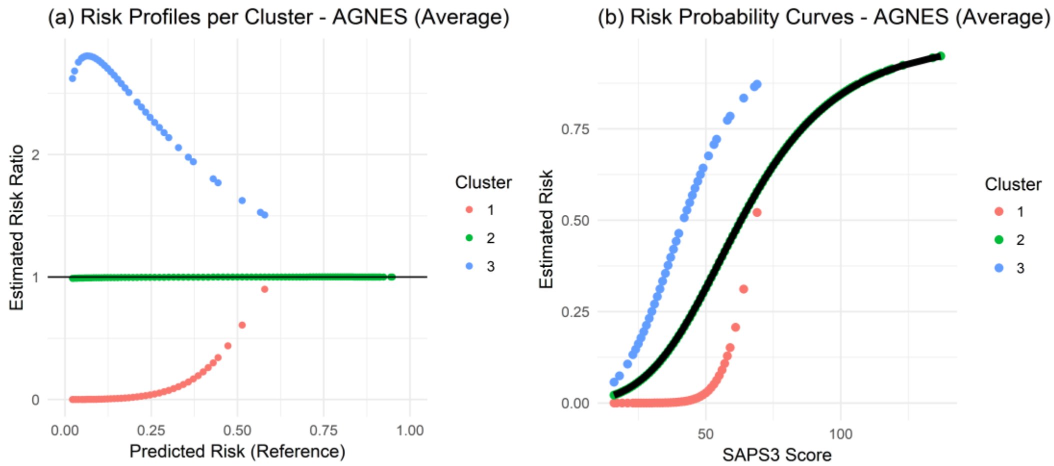

Figure 20 – Risk Ratios per cluster (AGNES-Average): (a) Risk Profile; (b) Predicted risks curves

The Risk Profile regarding the AGNES with Average linkage is shown in Figure 20 – (a). We can observe that Cluster 1 (lowest SMR) presents a crescent curve of Risk Ratios as the predicted risks increase. Therefore, the low severity cases have lower predicted risks compared to the reference. Conversely, Cluster 3 displays a high mortality risk in the low severity cases. The intermediate cluster risk ratio curve falls between the extreme clusters.

The curves converge to the reference risk ratio (1.00) as the predicted risks are close to 100%. This pattern is displayed in Figure 20 – (b), where we can verify that the predicted risks are below and above the reference line for the clusters with high and low SMR, respectively.

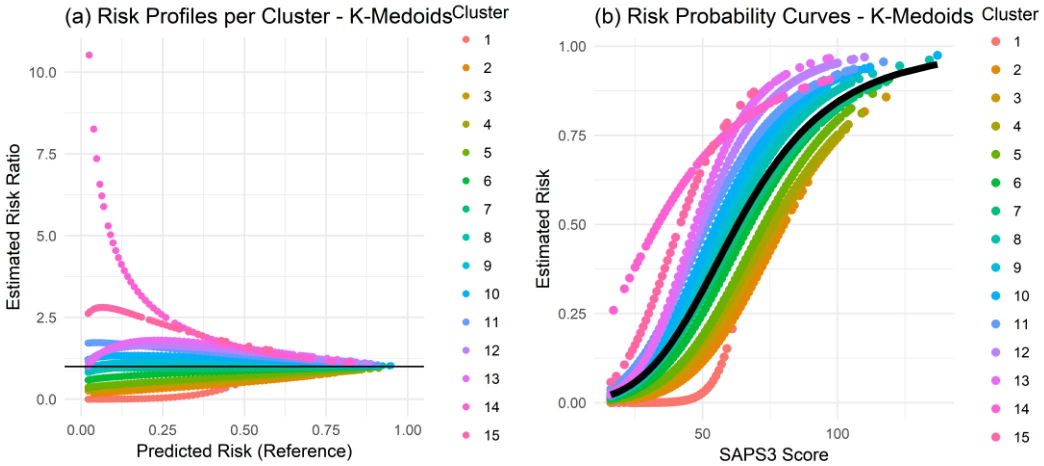

Figure 21 – Risk Ratios per cluster (K-Medoids): (a) Risk Profile; (b) Predicted risks curves

Similarly, we can observe the risk ratios for the K-Medoids clustering in Figure 21 – (a). K-Medoids has provided the largest number of clusters among the procedures, and we can notice a similar behavior of the extreme clusters as in the AGNES-Average. In this case, Cluster 14 presented the high value of risk ratios for the low severity cases, followed by the Cluster 15, which presents the highest SMR. Clusters 1 – 5 have SMR below 1.00 and their curves can be seen ordered under the reference line.

All the risk ratio curves converge to the reference risk ratio line. In Figure 21 – (b) we observed the same pattern on the predicted risks, being the curves over the reference line from the clusters with high SMR and the curves under the reference line present low SMR values. We verified that the extreme curves only present predicted risks from low severity cases, being one with higher predicted risks and the other with lower predicted risks compared to the reference.

Therefore, as we evidenced in the Clusters provided by the Rankability methodology, the best and worst clusters comprise of similar case-mixes, which as composed by low severity cases and, in the perspective of the risk ratios, the predicted risks are higher and lower than the reference, respectively. Furthermore, we observed that the risk ratios only present a large variation for lower predicted risks. The curves tend to converge to a risk ratio equal to one as the severity (predicted risks) increases.

5.5 Discussion on the SMR variability

Both clusters with best and worst performance present low severity cases and distinct SMR values. In this sense, we evaluated the relation between the SAPS3 median and the SMR values for the ICUs, as shown in Figure 22.

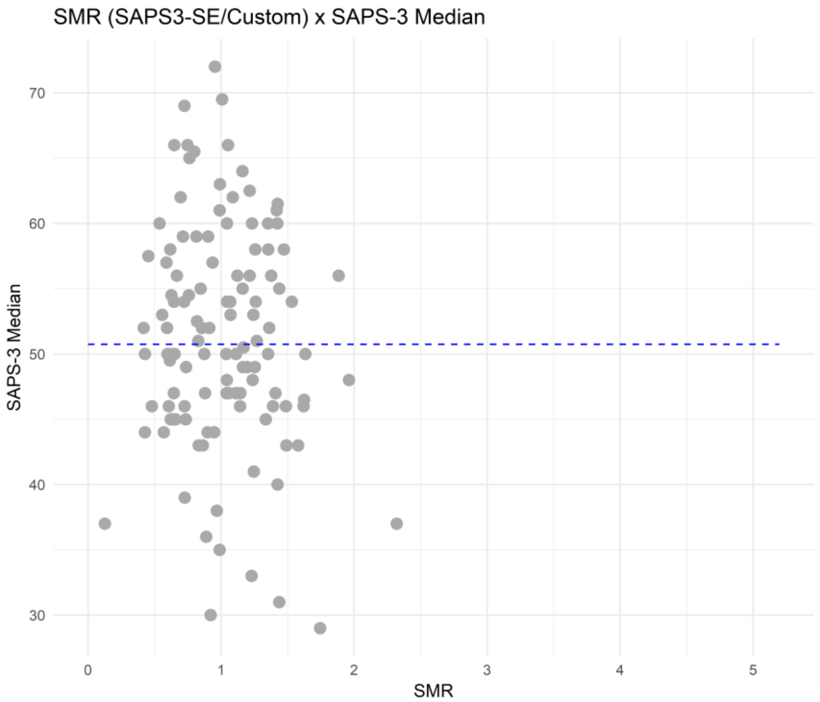

Figure 22 – SMR and SAPS-3 Median distribution

We observed that as the SAPS-3 Median increases (high severity cases) the SMR converges to one. Conversely, when we evaluate low severity scores, the values of SMR present a more significant dispersion, which includes the best and worst ICU performance. A triangular shape that represents the greatest spread in SMR values when the severity scores are smaller can be observed.

As previously discussed, the original SAPS3-SE provides biased predicted risks as it has a higher miscalibration on the extreme scores. However, we consider this equation to evaluate the variability of the SMR regarding the severity scores, since the customization tends to provide an effect adjusted by the study sample, compared to the original SAPS-3 sample estimator. The same plot (SMR x SAPS3 median) is considered with the original equation, as shown in Figure 23.

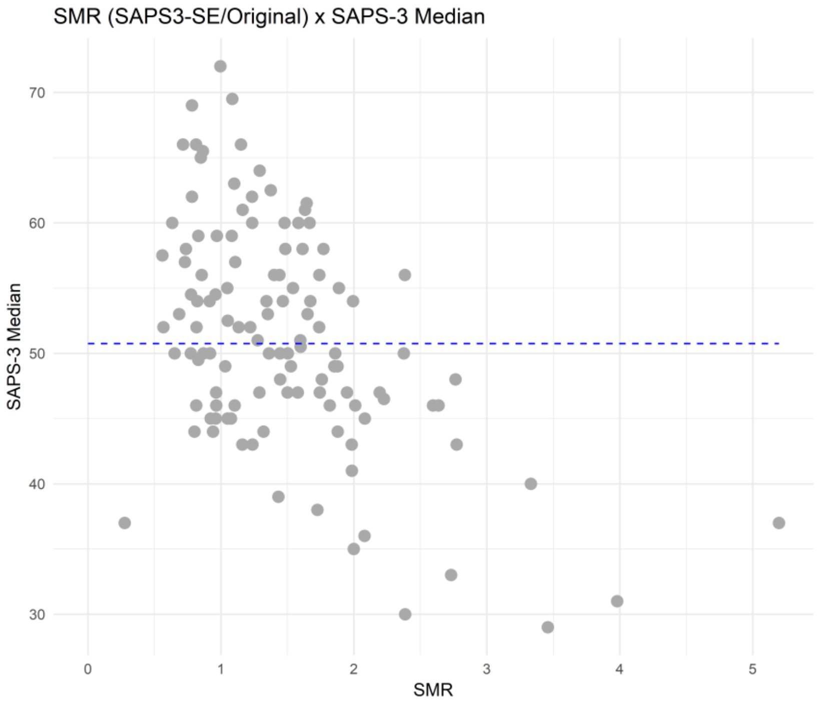

Figure 23 – SMR and SAPS-3 Median distribution

This spread pattern for ICUs with the median of severity scores lower than the overall median tends to provide a larger variability then ICUs with high-severity (and high-risk) case-mixes. We observed that, mathematically, the SMR never assume values below zero and the maximum value is directly related to the severity score. For this purpose, we evaluated two extreme cases in regarding the SMR range.

High Risk case-mix: If an ICU has a total of N Patients from high severity cases, the predicted mortality risks tend to be high, being the maximum expected deaths equal to the total number of patients. Therefore, the observed number of deaths cannot exceed the maximum expected number of deaths. Hence, the maximum value of SMR will not exceed 1.00, and any lower observed number of deaths results in a good SMR (below the reference).

Considering α the fraction of patients that really died from the total N patients admitted in the ICU, we have that the observed number of deaths $O = \alpha N$.

$$\lim_{E \to N} \frac{O}{E} = \frac{\alpha N}{N} = \alpha \tag{8}$$

Consequently, the SMR is equal to α, which ranges from [0,1].

Low-Risk case-mix: If an ICU has a total of N Patients from low severity cases, the predicted mortality risks tend to be low, being the minimum expected deaths equal zero. Therefore, any observed deaths can increase the SMR without limitations.

Considering α the fraction of patients that died from the total N patients admitted in the ICU, then the observed number of deaths $O = \alpha N$.

$$\lim_{E \to 0} \frac{O}{E} = \frac{\alpha N}{0} = \infty \tag{9}$$

Hence, for the same case-mix composed by high severity patients, there would hardly be an ICU with SMR higher than 1.00, which may suggest an incorrect interpretation of good performance. Conversely, for a low severity case-mix, the SMR may vary in the range [0, ∞), which may imply that the ICU with a large SMR may be considerably worse. However, its value must be compared relative to the maximum value (scale) that case-mix comprehends.

We identify that each case-mix has a range of predicted risks, with lower and higher possible values, that must be considered in comparison with the case-mix. To observe this pattern in a general approach, we simulated SMR values, using the SAPS3-SE original equations for different SAPS3 intervals in the study sample. We grouped patients for each 5% quantile SAPS3, which resulted in 20 groups or intervals of the severity score, resampled (with replacement) those patients in 500 replications and obtained the SMR values for each group. The result is shown in Figure 24.

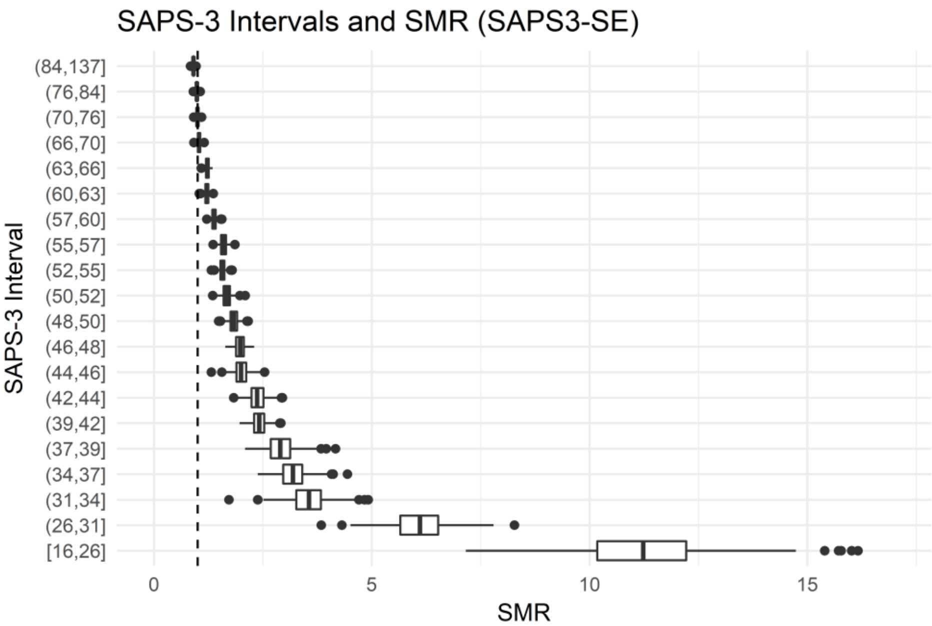

Figure 24 – SMR distribution per SAPS3 intervals

The separation for smaller quantiles provides less dispersion among SAPS3 values in the same interval, compared to the common deciles used in previous studies. We notice that the distribution of the SMR per severity group resembles the behavior in Figure 24. From a certain high SAPS3 value, the SMR presents lower spread and is less than one, while for low-risk cases the spread is extremely high.

Therefore, with this experiment, we verify that the SMR has a higher variability when the severity is low and, for each interval of SAPS3 (severity score) it has a defined range. It implies that there may be a biased analysis when comparing two or more ICUs with the same SMR and different case-mixes, since the indicator has limits that depend on the severity span of those ICUs.

The SAPS3 equation can provide an easy way of calculating the predicted risks using a reference, and the customization can reduce the bias in the estimation of extreme cases. However, it is necessary to understand the limitations when using the indicator to evaluate the performance regarding mortality among ICUs. In addition, better indicators (unbiased) for performance (either mortality or resource use) could be extremely helpful in the future to the study field.

6
Conclusion

The evaluation of performance in ICUs must consider important factors and outcomes to the patients and the stakeholders. The analysis of different indicators provides diagnostic on the status of an ICU care activities, and the detection of possible processes nonconformities that may increase inadequate care, mortality rates and higher costs. In this context, this study analyzed the performance on 116 ICUs with a total of 12,100 patients, considering the perspectives of mortality and resource use, as well as it evaluated the influences of institutional factors into those outcomes, combining different techniques for a more detailed performance evaluation compared to previous studies.

The study sample presented an overall lethality rate of 35% and most patients were admitted due to postoperative cares. As expected, patients who died had higher SAPS3 and SOFA scores than those who were discharged, also most of them were admitted due to high-risk conditions such as Sepsis. Regarding the ICUs, the sample comprised mostly of ICUs from public administration (47%), who assign most of their beds to the public healthcare system (75%), and from tertiary hospitals (78%). For this reason, the lethality rate was also high in those categories, which could evidence possible differences in care procedures.

To evaluate the performance and compare different ICUs, we calculated the SMR and SRU indicators for each unit. Firstly, we evidenced that the original SAPS3 equations (Standard and Central/South America) do not provide a good calibration with the sample's data and required a first-level customization to reduce the bias. Moreover, the SRU was the main resource use indicator considered in this study as it provides an interpretation similar to the SMR.

The efficiency matrix had evidence a strong correlation between the SMR and SRU indicators, and also that most ICUs are balanced between the most efficient and least efficient groups. We observed a small variability in the former group while the latter presented a larger range of SMR and SRU values.

ICU characteristics that presented higher lethality rate, variables related to "Hospital Administration" and "Bed-Assignment", showed a statistically significant difference among their groups regarding the mortality and resource use indicator. Private for-profit hospitals, and those who assign their beds to the private healthcare showed better performance, compared to public and SUS assignment. This may reflect the capability and better allocation of resources and investments within private hospitals, which can provide better care procedures.

Regarding the institutional factors, we evidenced that the performance indicators were significantly related to Organizational Procedures, Physical Infrastructure, Risk Management, Infection Prevention & Control routines, and Evaluation activities, in a univariate analysis. In the multivariate context, the combined variable of "Physical Infrastructure and Infection Prevention & Control" was noticed as a significant factor to both SMR and SRU. However, as those variables are estimated from the RDC-7 standard questionnaire, each category is limited to the number of questions provided and answered and the assigned weights. Hence, new approaches to identify factors or the estimate scores regarding adherences to best practices could improve this analysis.

To estimate performance groups, we performed clustering on the ICUs, using the SMR as the main variable. We evidenced that all clustering techniques resulted in unbalanced number of ICUs per cluster, being groups with extreme values of mortality (best and worst) comprised of single ICUs in most of the clustering procedures. In addition, we noticed that those extreme clusters presented a case-mix of low-risk patients compared to other clusters, which indicated a possible relation between the SMR variability and the severity span of a performance group.

Hence, we designed the profile risks of each cluster to analyze how the predicted risks of each cluster behave related to the whole sample. We verified that clusters with lower SMR values presented on average lower risk ratio for low-risk groups, while clusters with high SMR also demonstrated high-risk ratio values for the same risk group. For both cases, the risk ratio converges to one (predicted risk in the group = predicted risk in the sample). Therefore, we evidenced that the performance on mortality is related mainly to how the units or performance groups treat their low severity patients, which also indicated that there is a high variability on the risk ratio associated with the low severity group and the variability is reduced when the risks become high. This effect could also be presented in the SMR, thus incurring into bias.

In this sense, we also identified the variability effect in the SMR when relating the indicator of each unit with their respective case-mix (median SAPS3). This bias effect was intensified when we estimated the SMR using the predicted risks from the original SAPS3 equations. Hence, we evidenced that, depending on the severity risk, or the case-mix of the ICU, its SMR can only vary in a certain range: for high risk cases, the SMR tends to be at most equal to one, while for low-risk groups, the SMR tends to have a larger range. We attested this effect when simulating SMR values for SAPS3 intervals to analyze in a general perspective. Thus, for instance, comparing two ICUs with same SMR and different case-mix may not be completely reliable, since the SMR scale is different.

The analysis using SMR should be carefully used due to the bias effects that appear according to the case-mix. This pattern has been previously attested in other studies, however we presented the effect with an experimentation using the studied sample. Thus, this study makes the following contributions:

a) The analysis of performance in a group of ICUs considering three perspectives: the mortality, the resource use, and additional factors such as the characteristics and the institutional factors presented in the sample, while previous studies reported at most SMR and SRU values;
b) With the Rankability and Risk Profiles concepts we performed several common-used clustering techniques and combined the techniques to observe how performance groups are composed and the effects of grouping in the indicators. As far as we know, tests with different clustering techniques on the Rankability and its combination were not previously performed, and it support the methodology to evaluate ICUs;
c) The analysis on the SMR variability and its limitation: in the perspective of the ICUs and the overall severity span, we observed that the SMR has a variable range, which depends directly on the case-mix of a ICU, being smaller for high-risk patients, and larger for low-risk patients. As far as we know, this detailed analysis on the SMR was not reported previously, with experimental demonstration regarding the limitations of its variability.

We recall that as this database was obtained from a clinical trial study, the results obtained from the analysis of performance are evidences regarding this sample, that represents only a certain group of ICUs in Brazil.

6.1 Suggestions for future research

The analysis of performance is still an ongoing topic and new contributions can provide more adequate and applicable knowledge to improve decision making and guarantee quality of care in healthcare. In this sense, we also suggest future work to towards the following fields:

a) The continuous analysis and contributions regarding the benchmarking analysis

Different applications of benchmarking methodologies can provide different perspectives and improve processes within ICUs and hospitals. In addition, more representative databases should be created and evaluated to identity the performance on certain ICU groups, especially in public health perspective.

b) The development of new performance indicators

This study considered mainly two perspectives (mortality and resource use) and approached a third (institutional factors) to obtain more information and deep understanding on the benchmarking process. New indicators could be created to cover other domains, such as the impacts from the patient's care on its post-care and the perspective from the workers and how it influences the outcomes.

In addition, the continuous improvement of current indicators is also suggested. In this study, we observed that the original equations from the SAPS3 did not fit well on the sample, since they were estimated using a sample with particular characteristics, and that the SMR has limitations that could mislead interpretations when comparing ICUs.

c) The combination and proposal of new benchmark methods

The use of different methods can result in distinct perspective of analyzing a unit or a group of health facilities. The clustering procedure in the Rankability has a good potential to identify performance groups, as provided in this study, and the Risk Profiles could step into the SMR interpretation and demonstrate the mortality risks regarding each ICU, which assisted the identification on how the groups were performing regarding its case-mix. Hence, we suggest future research to consider combinations of different approaches for benchmarking to obtain a broad perspective on performance analysis.

7
References

AFESSA, B.; GAJIC, O.; KEEGAN, M. T. **S**everity of Illness and Organ Failure Assessment in Adult**. Intensive Care UnitsCritical Care Clinics**, 2007.

BROWN, S. E. S.; RATCLIFFE, S. J.; HALPERN, S. D. An Empirical Comparison of Key Statistical Attributes Among Potential ICU Quality Indicators*. **Critical Care Medicine**, v. 42, n. 8, p. 1821–1831, 2014.

CAPUZZO, M.; MORENO, R. P.; LE GALL, J.-R. Outcome prediction in critical care: the Simplified Acute Physiology Score models. **Current Opinion in Critical Care**, v. 14, n. 5, p. 485–490, 2008.

CAVALCANTI, A. B. et al. Effect of a Quality Improvement Intervention With Daily Round Checklists, Goal Setting, and Clinician Prompting on Mortality of Critically Ill Patients: A Randomized Clinical Trial. **Jama**, v. 315, n. 14, p. 1480–1490, 2016.

FINAZZI, S. et al. Calibration belt for quality-of-care assessment based on dichotomous outcomes. **PLoS ONE**, v. 6, n. 2, 2011.

GARLAND, A. Improving the ICU: Part 2. **Chest**, v. 127, n. 6, p. 2165–2179, 2005.

GOLDSTEIN, H.; SPIEGELHALTER, D. J. League tables and their limitations: Statistical issues in comparisons of institutional performance. **Journal of the Royal Statistical Society. Series A (Statistics in Society)**, v. 159, n. 3, p. 385–443, 1996.

HIGGINS, T. L. et al. Assessing contemporary intensive care unit outcome: An updated Mortality Probability Admission Model (MPM0-III)*. **Critical Care Medicine**, v. 35, n. 3, p. 827–835, 2007.

HUBERT COMLAVI: CASTRA, LAURENT: SOUBRIER, STEPHANE: VILHELM, CHRISTIAN: RAVAUX, PIERRE: LEMDANI, MOHAMED: DUROCHER, ALAIN: SAULNIER, FABIENNE, H. G. Methodological approach for the evaluation of the performances of medical intensive care units. **Journal of critical care**, v. 22, n. 3, p. 184–190, 2007.

KAUFMAN, L.; ROUSSEEUW, P. J. **Finding Groups in Ordinal Data. An Introduction to Cluster Analysis**. [s.l: s.n.].

KEEGAN, M. T.; SOARES, M. What every intensivist should know about prognostic scoring systems and risk-adjusted mortality. **Revista Brasileira de Terapia Intensiva**, 2016.

KHANMOHAMMADI, S.; ADIBEIG, N.; SHANEHBANDY, S. An improved overlapping k-means clustering method for medical applications. **Expert Systems with Applications**, v. 67, p. 12–18, 2017.

KRAMER, A. A. A Flock of Birds, a Cluster of ICUs. **Critical Care Medicine**, v. 44, n. 5, p. 1016–7, 2016.

KRAMER, A. A.; ZIMMERMAN, J. E. The relationship between hospital and intensive care unit length of stay*. **Critical Care Medicine**, v. 39, n. 5, p. 1015–1022, 2011.

LONE, N. I. et al. Five-year mortality and hospital costs associated with surviving intensive care. **American Journal of Respiratory and Critical Care Medicine**, v. 194, n. 2, p. 198–208, 2016.

METNITZ, B. et al. Austrian validation and customization of the SAPS 3 Admission Score. **Intensive Care Medicine**, v. 35, n. 4, p. 616–622, 2009.

METNITZ, P. G. H. et al. SAPS 3-From evaluation of the patient to evaluation of the intensive care unit. Part 1: Objectives, methods and cohort description. **Intensive Care Medicine**, v. 31, n. 10, p. 1336–1344, 2005.

MINISTÉRIO DA SAÚDE. Resolução-RDC n 7, de 24 de fevereiro de 2010. **Diário Oficial da União**, p. 11, 2010.

MIRANDA, D. R.; DE RIJK, A.; SCHAUFELI, W. Simplified therapeutic intervention scoring system: The TISS-28 items--results from a multicenter study. **Critical Care Medicine**, v. 24, n. 1, p. 64–73, 1996.

MORALEZ, G. M. et al. External validation of SAPS 3 and MPM0-III scores in 48,816 patients from 72 Brazilian ICUs. **Annals of Intensive Care**, v. 7, n. 1, p. 53, 2017.

MORENO, R. P. et al. 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**, v. 31, n. 10, p. 1345–1355, 2005.

MORENO, R. P. et al. Characterizing the risk profiles of intensive care units. **Intensive Care Medicine**, v. 36, n. 7, p. 1207–1212, 2010.

NATHANSON, B. H. et al. A revised method to assess intensive care unit clinical performance and resource utilization*. **Critical Care Medicine**, v. 35, n. 8, p. 1853–1862, 2007.

NISKANEN, M.; REINIKAINEN, M.; PETTILÄ, V. Case-mix-adjusted length of stay and mortality in 23 Finnish ICUs. **Intensive Care Medicine**, v. 35, n. 6, p. 1060–1067, 2009.

POOLE, D. et al. Comparison between SAPS II and SAPS 3 in predicting hospital mortality in a cohort of 103 Italian ICUs. Is new always better? **Intensive Care Medicine**, v. 38, n. 8, p. 1280–1288, 2012.

POUW, M. E. et al. Hospital Standardized Mortality Ratio: Consequences of Adjusting Hospital Mortality with Indirect Standardization. **PLoS ONE**, v. 8, n. 4, 2013.

RAPOPORT, J. et al. A method for assessing the clinical performance and cost-effectiveness of intensive care units: a multicenter inception cohort study. **Critical Care Medicine**, v. 22, n. 9, p. 1385–1391, 1994.

RAY, B. et al. Quality indicators for ICU: ISCCM guidelines for ICUs in India. **Indian journal of critical care medicine : peer-reviewed, official publication of Indian Society of Critical Care Medicine**, v. 13, n. 4, p. 173–206, 2009.

REIS MIRANDA, D.; MORENO, R.; IAPICHINO, G. Nine equivalents of nursing manpower use score (NEMS). **Intensive Care Medicine**, v. 23, n. 7, p. 760–765, 1997.

REYNOLDS, A. P. et al. Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. **Journal of Mathematical Modelling and Algorithms**, v. 5, n. 4, p. 475–504, 2006.

ROTHEN, H. U. et al. Variability in outcome and resource use in intensive care units. **Intensive Care Medicine**, v. 33, n. 8, p. 1329–1336, 2007.

ROTHEN, H. U.; TAKALA, J. Can outcome prediction data change patient outcomes and organizational outcomes? **Current Opinion in Critical Care**, v. 14, n. 5, p. 513–519, 2008.

SALLUH, J. I. F.; SOARES, M.; KEEGAN, M. T. Understanding intensive care unit benchmarking. **Intensive Care Medicine**, v. 43, n. 11, p. 1703–1707, 2017.

SHWARTZ, M.; YOUNG, D. W.; SIEGRIST, R. The ratio of costs to chargees: how good a basis for estimating costs? **Inquiry**, v. 32, n. 4, p. 476–481, 1995.

SIEGEL, T. et al. Prospective assessment of standardized mortality ratio (SMR) as a measure of quality of care in intensive care unit--a single-centre study. **Anaesthesiology intensive therapy**, v. 47, n. 4, p. 328–332, 2015.

SILVA JUNIOR, J. M. et al. Applicability of the simplified acute physiology score (SAPS 3) in Brazilian hospitals. **Revista brasileira de anestesiologia**, v. 60, n. 1, p. 20–31, 2010.

SOARES, M. et al. Organizational characteristics, outcomes, and resource use in 78 Brazilian intensive care units: the ORCHESTRA study. **Intensive Care Medicine**, v. 41, n. 12, p. 2149–2160, 2015.

VAN DISHOECK, A.-M. et al. Random variation and rankability of hospitals using outcome indicators. **BMJ Quality & Safety**, v. 20, n. 10, p. 869–874, 2011.

VERBURG, I. W. M. et al. Individual and Clustered Rankability of ICUs According to Case-Mix–Adjusted Mortality*. **Critical Care Medicine**, v. 44, n. 5, p. 901–909, 2016.

VERGARA, S. Projetos e relatórios de pesquisa em administração. **Revista de Administração Contemporânea**, n. 10, p. 221–221, 2009.

WOODHOUSE, D. et al. Will benchmarking ICUs improve outcome? **Current Opinion in Critical Care**, v. 15, n. 5, p. 450–455, 2009.

ZIMMERMAN, J. E. et al. Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients*. **Critical Care Medicine**, v. 34, n. 5, p. 1297–1310, 2006.

APPENDIX I – RDC-7 Standard Categories and Items

Table 16 – RDC-7 Standard for best practices in ICUs

Category	Sub-items	Choices (weights)
Organization	Sepsis Sedation Analgesia Ventilatory Weaning Prevention of Ventilator-Associated Pneumonia Prevention of Central-Line Associated Bloodstream rate	(0) No; (1) Yes
Physical Infrastructure	Different Rooms for Adult, Pediatric or Neonatal ICUs? Privacy (at least separation by curtains between beds) Is there an insulation bed (at least one for each ten beds)?	(0) No; (1) Yes
Human Resources	Does the responsible technician hold a intensive care specialist title? Is the nursing coordinator a specialist in intensive care or in other specialty related to the assistance of severe patients? Is the physiotherapist coordinator a specialist in intensive care or in other specialty related to the assistance of severe patients?	(0) No/Not applicable; (1) Yes
	Is there a routine phyisician for each ten beds in morning or evening period?	(0) No; (0.5) Partially; (1) Yes
	Is there an exclusive phyisician on duty for each ten beds or fraction in every period? Does the unit have an exclusive nurse for each ten beds or fraction in every period?	(0) No; (1) Yes
	Does the unit have an exclusive physiotherapist for each ten beds or fraction in every period?	(0) No; (0.5) Partially; (1) Yes
	Does the unit have a nursing technician for each two beds in each period? Does the unit have at least one exclusive administrative assistant?	(0) No; (1) Yes

Table 17 – RDC-7 Standard for best practices in ICUs (cont.)

Category	Sub-items	Choices (weights)
Assistive Resources	Nutritional Assitance (with enteral and parenteral nutrition) Nephrologic Assistance (including hemodialysis) Hemotherapic Assistance Infectoloty Clinical Assistance General Sugery Assistance Clinical Laboratory service (including microbiology and hemogasometry) Mobile Radiography service Portable Ultrasonography service Digestive Endoscopy (upper and lower) service Fiberoptic bronchoscopy service Surgical Center Echocardiography service Cardiovascular Surgery Neurologic Surgery Interventional Radiology Computer Tomography Confirmatory tests for brain blood flow	(0) No; (1) Yes
Work Procedures	Orientation to relatives: Is there at least one daily period to contact with relatives?	(0) No; (1) Yes
	Vistis: How many daily visits of at least 30 minutes are permi ted?	(0) None; (0.5) One; (1) Full-time
Patient Transportation	Are all severe patients always moved with continuous assistance of at least one physician and one nurse? Are patients sistematicaly moved with multiparameter monitor, mechanical ventilation (for intubated patients) andOs pacientes são sistematicamente transportados com monitor multiparamétrico, ventilador ecânico (para os pacientes intubados) e com transportation kit?	(0) No; (1) Yes
Risk Management	Is there a routine of registering adverse events? Is there a person resposible to manage adverse events?	(0) No; (1) Yes
Infection Prevention and Control	Does the Hospital Infection Control Committee actively research infections related to invasive devices, multiresistant and other important microorganisms for clinical epidemiology? Does the Hospital Infection Control Committee report periodcally (at least each 3 months) the results from infection survaillance and sensitive profile of microorganisms to the intensive care staff? Is there alcoholic prepation for hand cleaning available at the unit's entrance and between beds?	(0) No; (1) Yes

Table 18 – RDC-7 Standard for best practices in ICUs (end)

Category	Sub-items	Choices (weights)
Evaluation	Absolute mortality Rate	(0) No; (0.5) Partially; (1) Yes
	Mortality rate estimated from the severity scores	(0) No; (1) Yes
	Average ICU stay time	
	24 hours readmission rate	
	Ventilator-Associated Pneumonia (VAP) incidence density	
	Mechanical Ventilation (MV) utilization rate	
	Central-Line Associated Bloodstream infection rate density	
	Central-Line Catheter utilization rate	
	Urinary Tract infection rate density	
Material Resources	Manual resuscitator with reservatory and facial mask, one per bed, with operational reserve of one unit per two beds	(0) No; (1) Yes
	Four continouos and controlled infusion equipments ("infusion pump") with operational reserve of one equipment for each three beds (4.3 infusion pumps per beds)	
	Multiparameter monitoring (with at least respiractory frequency, pulse oximeter, cardioscopy, heart rate, temperature, non-invasive blood pressure)	
	"Cuffometer"	
	Mechanical Ventilator: one unit for each two beds, with operational reserve of one unit for each 5 beds, and two complete circuits per equipment (0.7 ventilaters per bed)	
	Non-invasive mechanical ventilator: one for each ten ebds, when the microprocessed mechanical ventilator cannot provide non-invasive vetilation (0.5 per bed)	
	Portable Electrocardiogram: one equipment for each ten beds	
	Defibrilator/cardioverter kit with medicines and resources for emergencies: one for each five beds or fraction	
	Temporary cardiac pacemkaer, electrodes and generator: one equipment for each ten beds	
	Refrigerator, with internal temperature of 2 to 8°C, exclusive for the storage of medicines, with temperature control	

APPENDIX II – Description of Clustering techniques

We describe briefly the assumptions and procedures used in the clustering techniques approached in this study as defined in Khanmohammadi et al. (2017) Reynolds et al. (2006), and Kaufman and Rousseeuw (2005) which can be found in Table 17.

Table 19 – Descriptive Information of Clustering Techniques. Source: Based on and Reynolds et al. (2006)

Type	Technique	Description
Hierarchical Clustering	Agglomerative Nesting (ANGES)	Agglomerative clustering is a technique in which single nodes are merged together until the whole dataset is a single cluster. This procedure builds a dendrogram in which a hierarchy is computed regarding the number of clusters. The merge procedure depends on the linkage function used to estimate de distance between two clusters and the criteria to combine two clusters by calculating their dissimilarity. In this study, we evaluated four linkage functions used in AGNES (KHANMOHAMMADI et al, 2017): - Single linkage: considers the minimum distance among the components of two clusters; - Complete linkage: conversely, it merges two clusters that have the maximum distance between them; - Average linkage: merges two clusters with the lowest average distance; - Ward's distance: it considers the minimum within-cluster variance.
	Divisive Analysis (DIANA)	In contrary to AGNES, it begins with the whole dataset as a single cluster and iteratively divides itself into clusters from it until all the data points are clusters considering the minimum distance among the data points within a cluster. It also computes a hierarchy in which one can choose the best number of clusters (Kaufman and Rousseeuw, 2005).
Partitional Methods	K-Means	An initial number of k clusters must be defined. It selects k points randomly as centroids and assigns clusters the data points closer to the respective centroid. In K-Means, the centroid corresponds to the mean of the coordinates from the points within the same cluster. Hence, in each iteration the centroids are recalculated, and data points are re-assigned to clusters to minimize the total average distance to the centroid (REYNOLDS et al., 2006).
	K-Medoids	An initial number of k clusters must be defined. It selects k data points as centroids randomly and assigns clusters the data points closer to the respective centroid. In K-Medoids, the centroid are data points. Hence, in each iteration the centroids are recalculated, and data points are re-assigned to clusters to minimize the total average distance to the centroid (REYNOLDS et al., 2006).

APPENDIX III – Efficiency Matrix and Institutional Variables

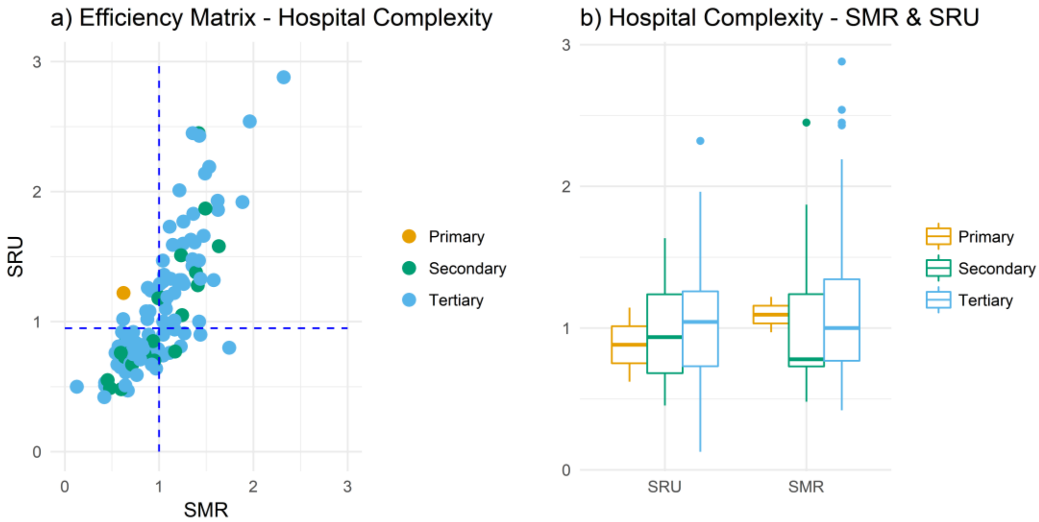

Figure 25 – Hospital Complexity: (a) Efficiency Matrix; (b) SMR and SRU Boxplot

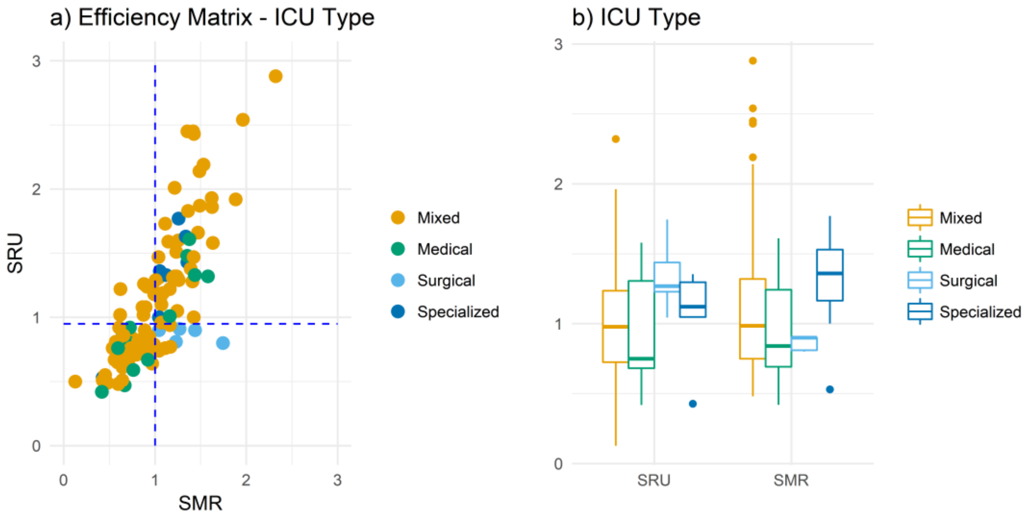

Figure 26 – ICU Type: (a) Efficiency Matrix; (b) SMR and SRU Boxplot

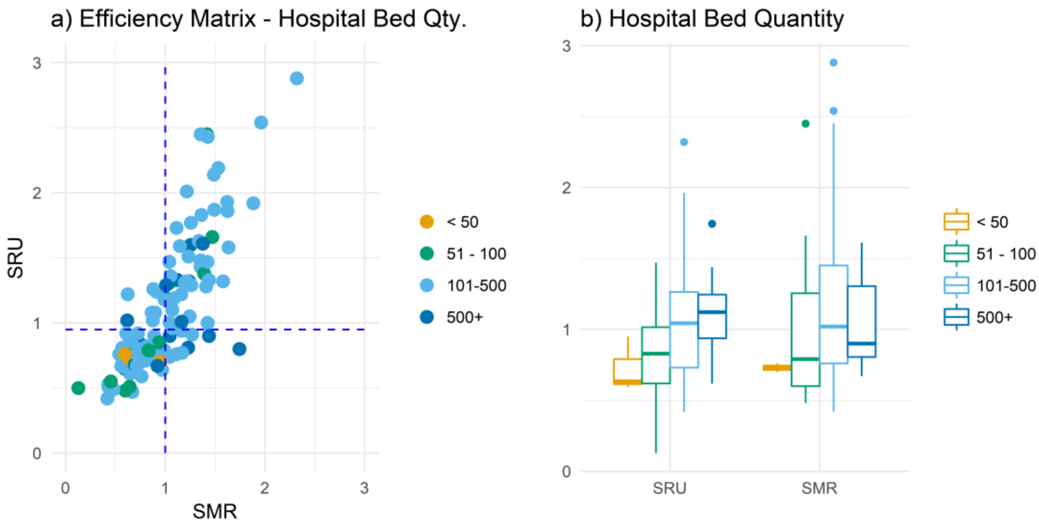

Figure 27 – Hospital Bed Quantity: (a) Efficiency Matrix; (b) SMR and SRU Boxplot

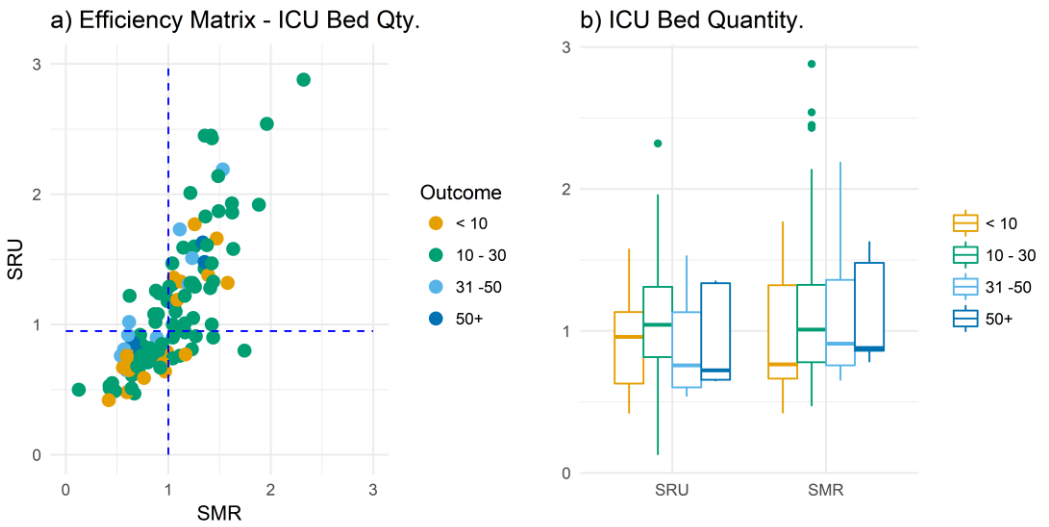

Figure 28 – ICU Bed Quantity: (a) Efficiency Matrix; (b) SMR and SRU Boxplot

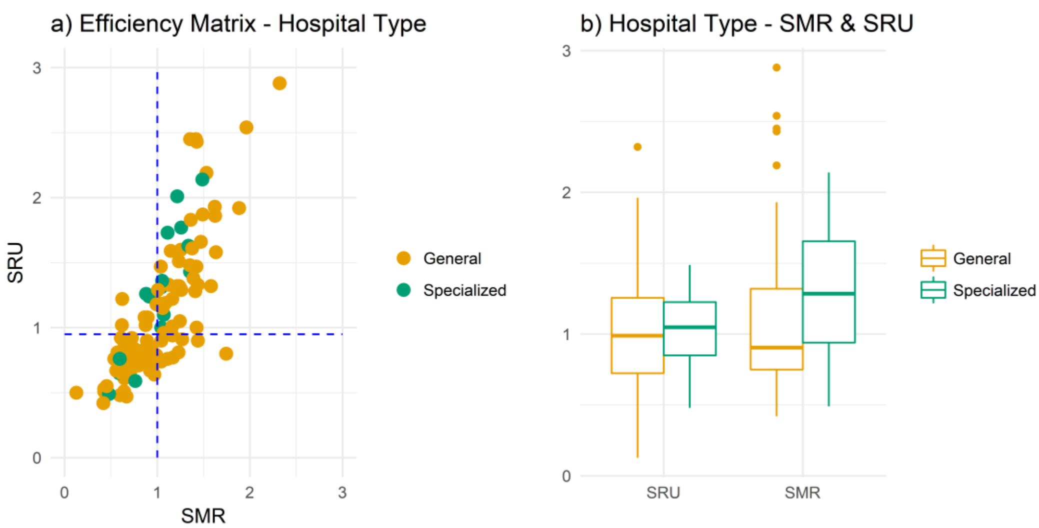

Figure 29 – Hospital Type: (a) Efficiency Matrix; (b) SMR and SRU Boxplot

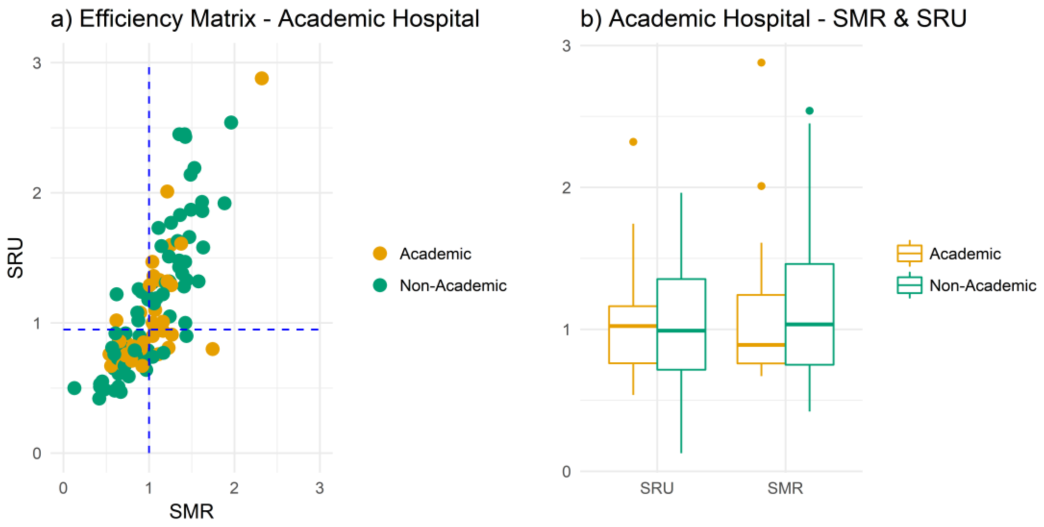

Figure 30 – Academic Hospital: (a) Efficiency Matrix; (b) SMR and SRU Boxplot

APPENDIX IV – Rankability clustering procedure

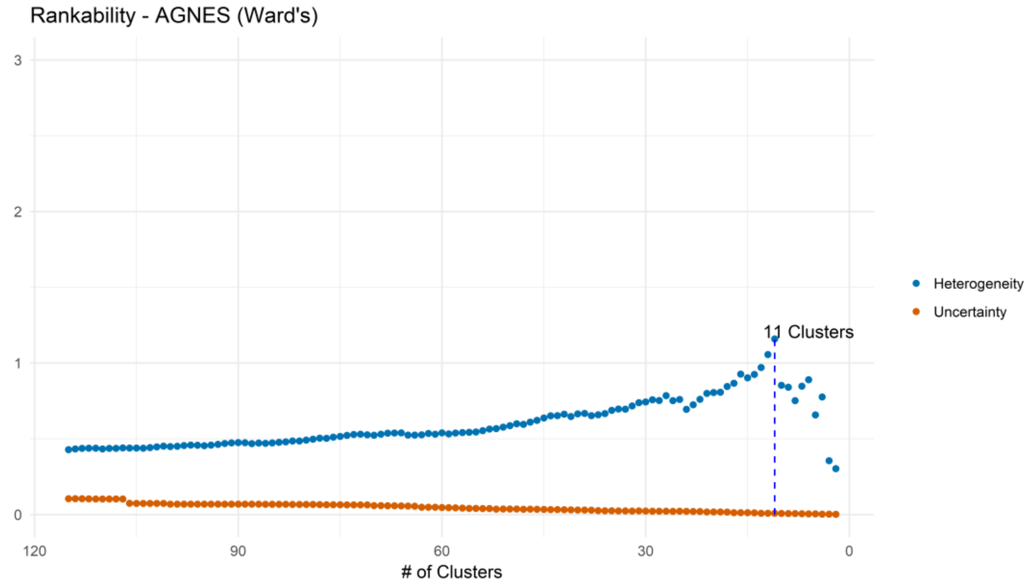

Figure 31 – Iterative Heterogeneity and Uncertainty - AGNES with Ward’s Distance linkage

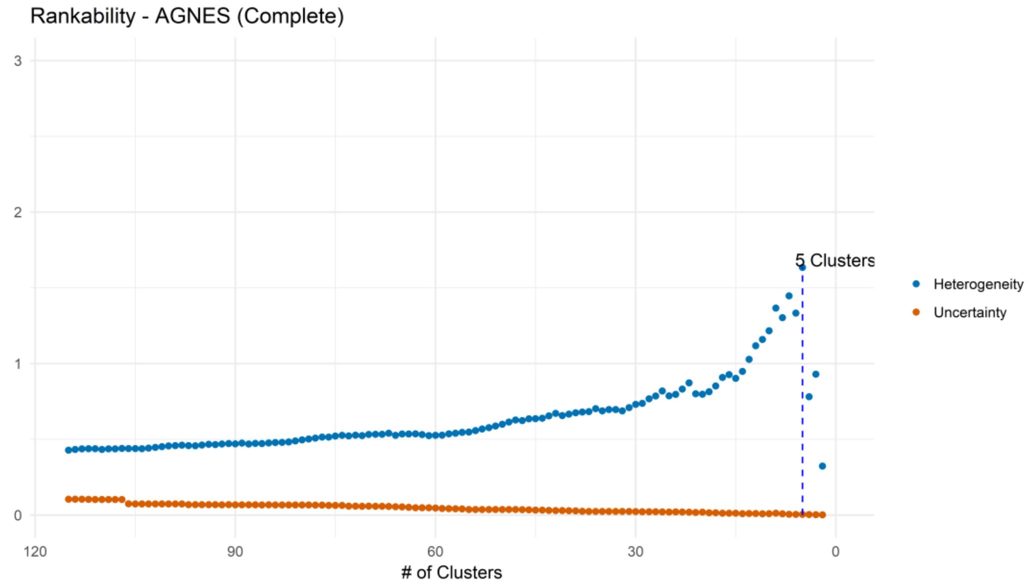

Figure 32 – Iterative Heterogeneity and Uncertainty - AGNES with Complete linkage

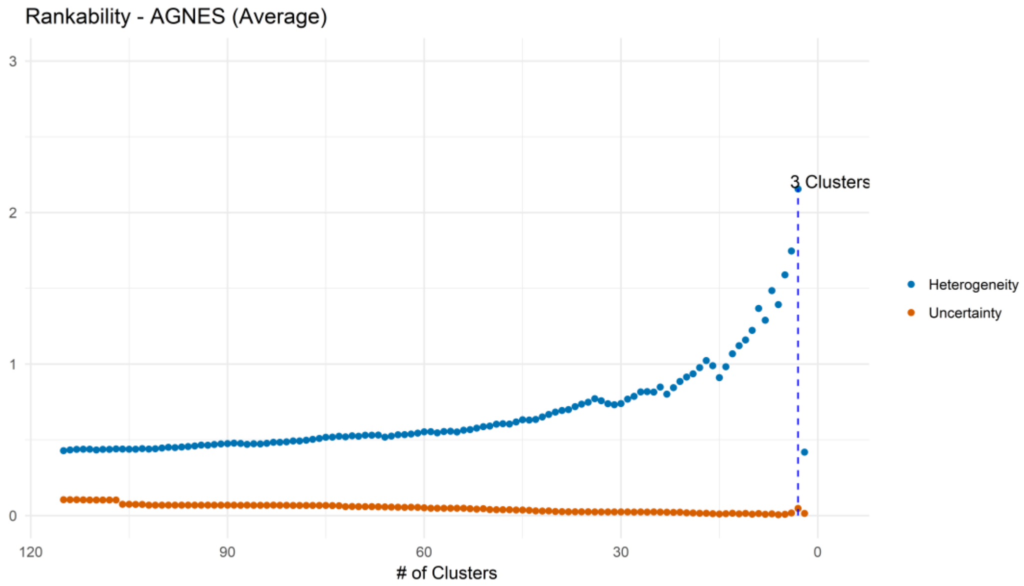

Figure 33 – Iterative Heterogeneity and Uncertainty - AGNES with Average linkage

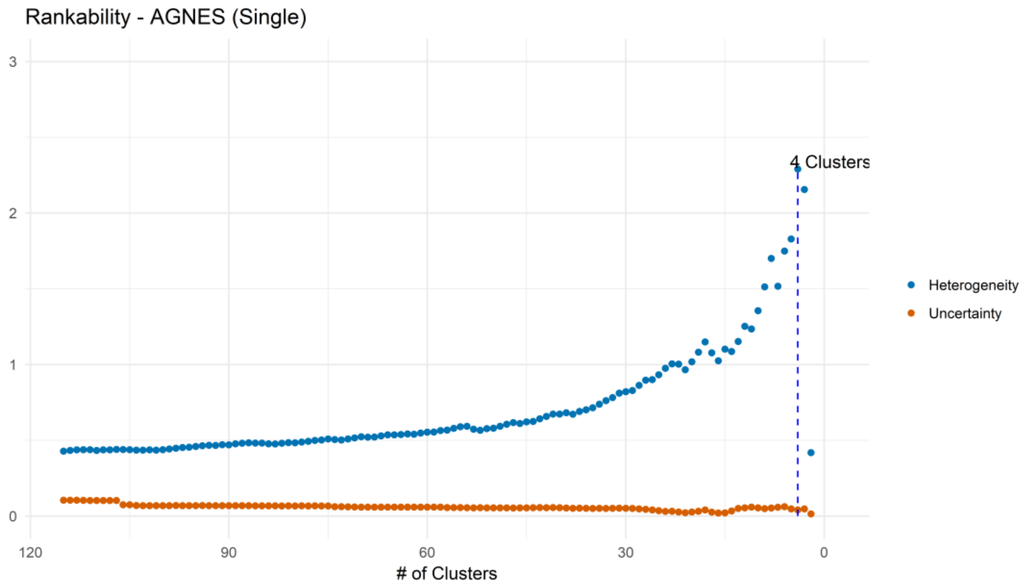

Figure 34 – Iterative Heterogeneity and Uncertainty - AGNES with Single linkage

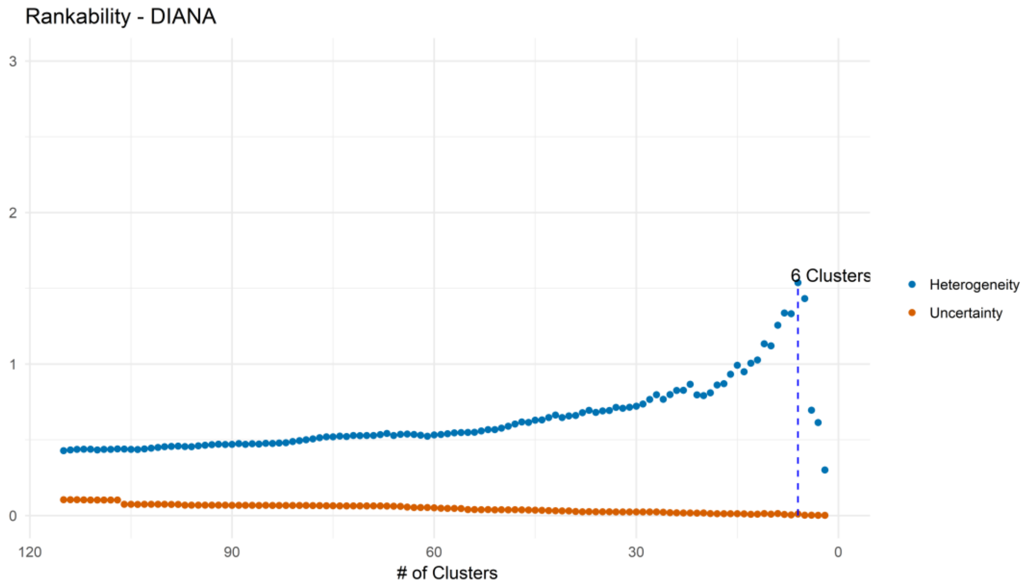

Figure 35 – Iterative Heterogeneity and Uncertainty - DIANA

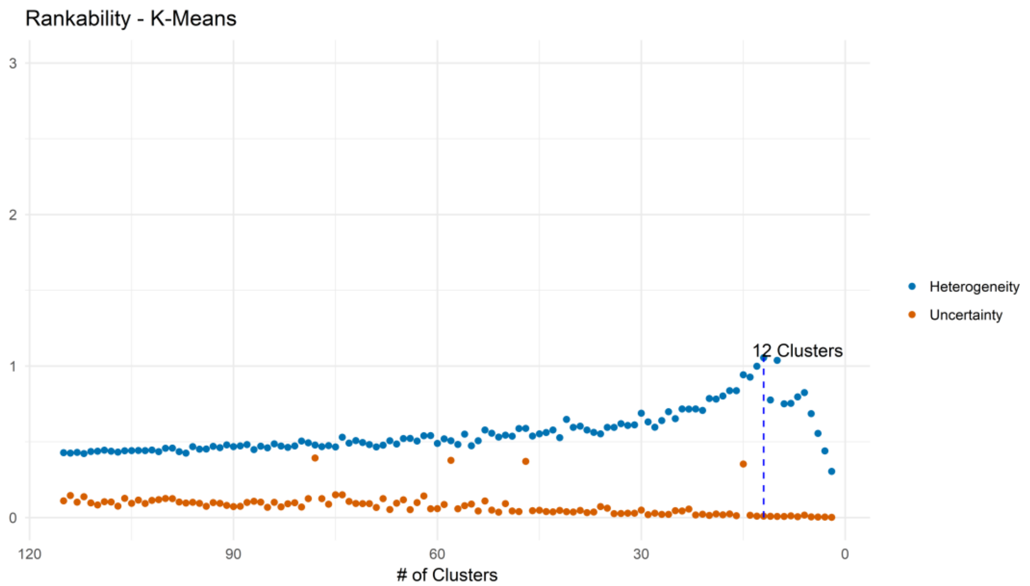

Figure 36 – Iterative Heterogeneity and Uncertainty - K-Means

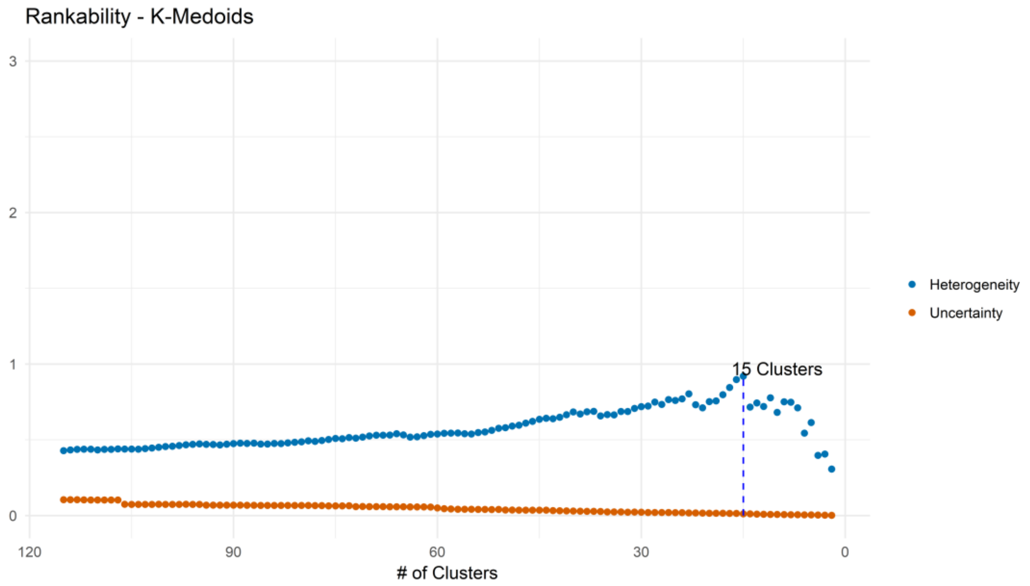

Figure 37 – Iterative Heterogeneity and Uncertainty - K-Medoids

APPENDIX V – Descriptive Information per Clustering

Table 20 – Descriptive Information - AGNES with Ward's Distance linkage

AGNES (Ward)						
Cluster	# Patients	Cluster SMR	SAPS3 Median, IQR		Lethality	# ICUs
1	133	0.13	37	30-44	2.26%	1
2	537	0.44	48	40-57	13.78%	5
3	1934	0.62	51	41-63	22.03%	19
4	1199	0.73	55	43-69	29.52%	11
5	2042	0.88	49	37-62	29.19%	19
6	1470	1.04	54	42-68	40.88%	14
7	2241	1.20	51	40-62	42.08%	21
8	1690	1.41	50	39-63	48.70%	17
9	586	1.62	45	33-56	44.03%	6
10	165	1.92	50	40-64	67.27%	2
11	103	2.32	37	30-43.5	39.81%	1

Table 21 – Descriptive Information - AGNES with Complete linkage

AGNES (Complete)						
Cluster	# Patients	Cluster SMR	SAPS3 Median, IQR		Lethality	# ICUs
1	133	0.13	37	30-44	2.26%	1
2	3670	0.64	51	41-64	23.27%	35
3	5753	1.05	51	40-63	37.20%	54
4	2276	1.45	48	37-61	47.50%	23
5	268	2.02	43	35-56	56.72%	3

Table 22 – Descriptive Information - AGNES with Average linkage

AGNES (Average)						
Cluster	# Patients	Cluster SMR	SAPS3 Median, IQR		Lethality	# ICUs
1	133	0.13	37	30-44	2.26%	1
2	11864	1.00	51	40-63	35.28%	114
3	103	2.32	37	30-43.5	39.81%	1

Table 23 – Descriptive Information - AGNES with Single linkage

AGNES (Single)						
Cluster	# Patients	Cluster SMR	SAPS3 Median, IQR		Lethality	# ICUs
1	133	0.13	37	30-44	2.26%	1
2	11699	0.99	51	40-63	34.83%	112
3	165	1.92	50	40-64	67.27%	2
4	103	2.32	37	30-43.5	39.81%	1

Table 24 – Descriptive Information - DIANA

DIANA						
Cluster	# Patients	Cluster SMR	SAPS3 Median, IQR		Lethality	# ICUs
1	133	0.13	37	30-44	2.26%	1
2	3670	0.64	51	41-64	23.27%	35
3	3211	0.94	50	39-64	33.23%	30
4	4715	1.30	50	40-62	45.17%	46
5	268	1.89	41	31-56	50.37%	3
6	103	2.32	37	30-43.5	39.81%	1

Table 25 – Descriptive Information from clusters in K-Means

K-Means						
Cluster	# Patients	Cluster SMR	SAPS3 Median, IQR		Lethality	# ICUs
1	133	0.13	37	30-44	2.26%	1
2	537	0.44	48	40-57	13.78%	5
3	1934	0.62	51	41-63	22.03%	19
4	1199	0.73	55	43-69	29.52%	11
5	1371	0.85	50	39-61	28.67%	13
6	1076	0.96	49	36-67	34.29%	10
7	1575	1.08	51	41.5-63	39.11%	15
8	1731	1.22	52	40-64	44.02%	16
9	1494	1.40	51	40-64	49.53%	15
10	679	1.57	47	36.5-58	46.69%	7
11	268	1.89	41	31-56	50.37%	3
12	103	2.32	37	30-43.5	39.81%	1

Table 26 – Descriptive Information from clusters in K-Medoids

K-Medoids						
Cluster	# Patients	Cluster SMR	SAPS3 Median, IQR		Lethality	# ICUs
1	133	0.13	37	30-44	2.26%	1
2	537	0.44	48	40-57	13.78%	5
3	860	0.58	51	41-64	21.28%	9
4	1074	0.64	50	40-62	22.63%	10
5	1199	0.73	55	43-69	29.52%	11
6	1371	0.85	50	39-61	28.67%	13
7	1002	0.96	48	35-64	32.83%	9
8	1139	1.05	54	43-67	41.70%	11
9	954	1.15	50.5	41-62	40.15%	9
10	1287	1.24	51	39-62	43.51%	12
11	651	1.36	52	41-64	50.23%	7
12	1039	1.44	48	36-62	47.74%	10
13	586	1.62	45	33-56	44.03%	6
14	165	1.92	50	40-64	67.27%	2
15	103	2.32	37	30-43.5	39.81%	1

APPENDIX VI – Risk Profiles per Clustering Techniques

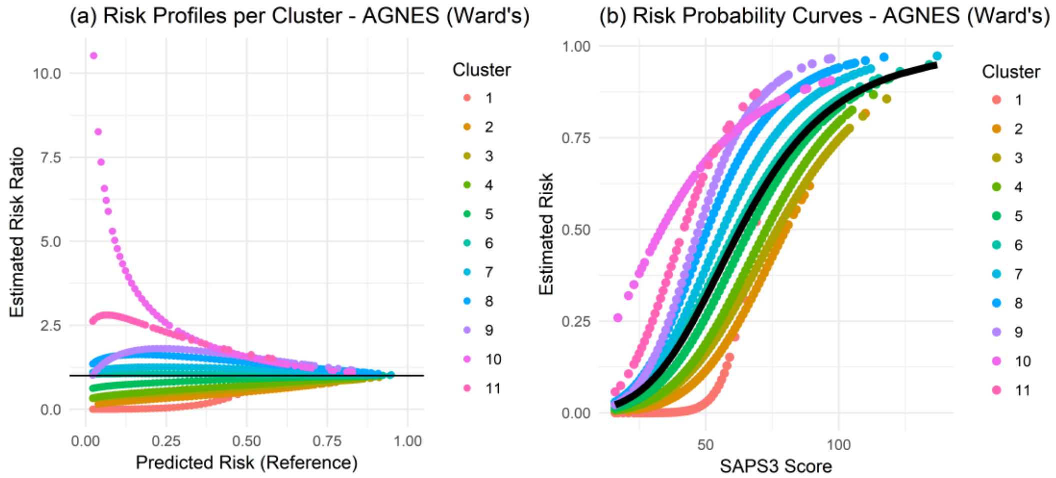

Figure 38 – Risk Ratios per cluster (AGNES – Ward’s): (a) Risk Profile; (b) Predicted risks curves

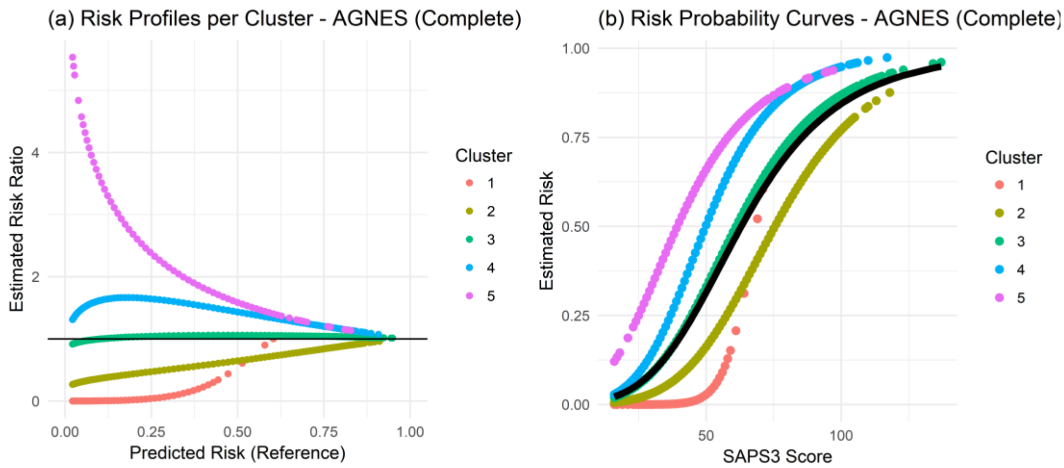

Figure 39 – Risk Ratios per cluster (AGNES – Complete): (a) Risk Profile; (b) Predicted risks curves

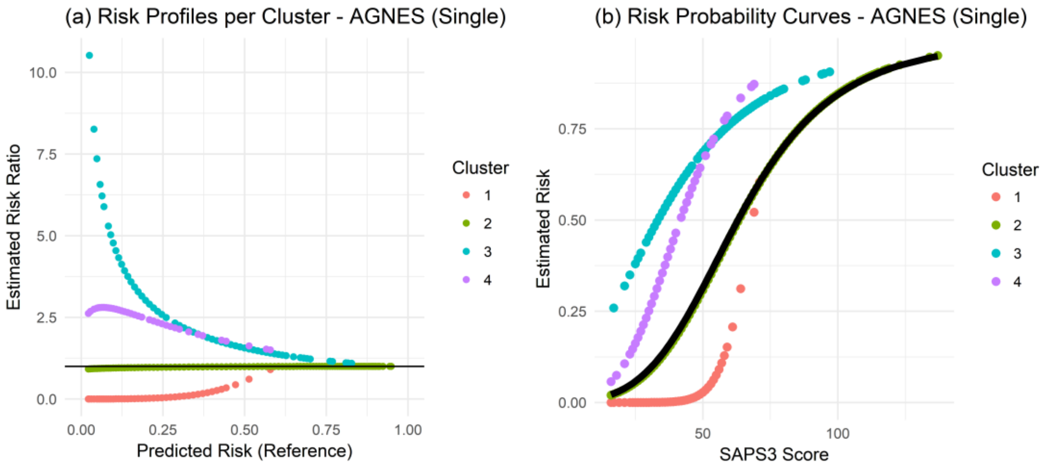

Figure 40 – Risk Ratios per cluster (AGNES – Single): (a) Risk Profile; (b) Predicted risks curves

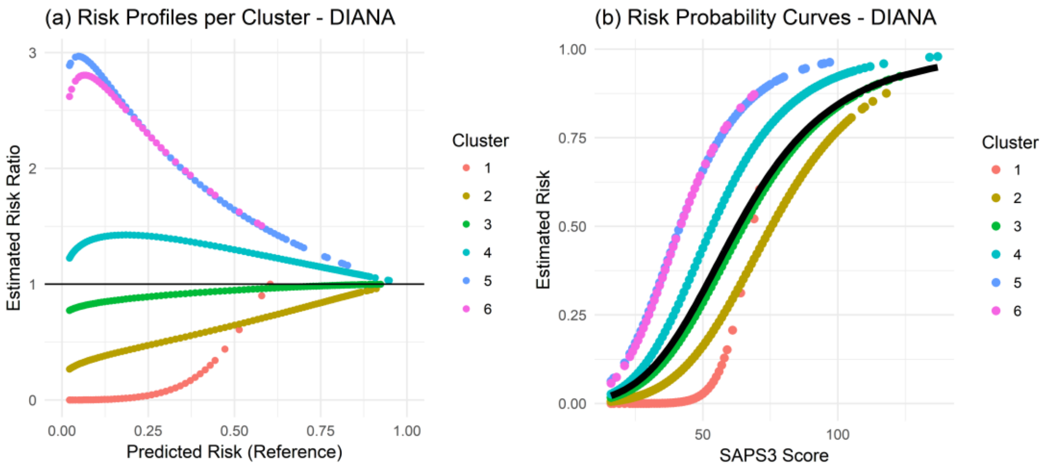

Figure 41 – Risk Ratios per cluster (DIANA): (a) Risk Profile; (b) Predicted risks curves

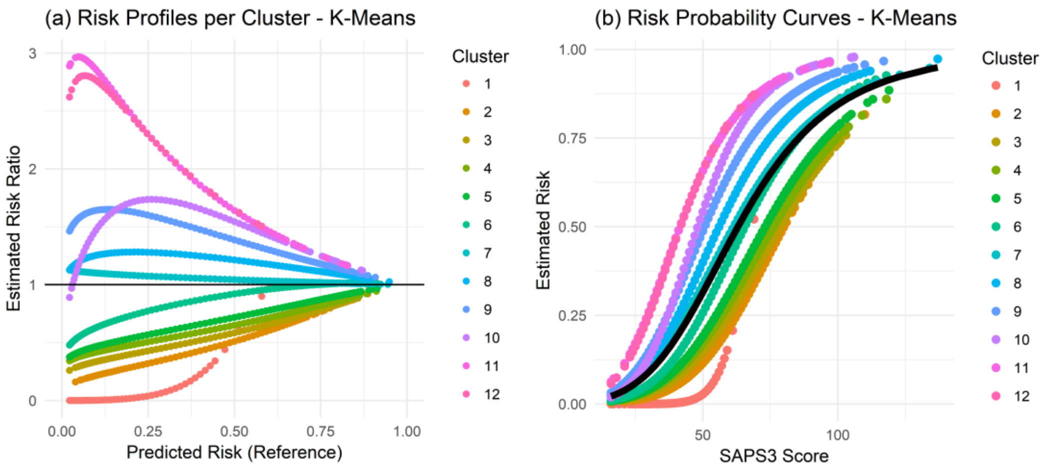

Figure 42 – Risk Ratios per cluster (K-Means): (a) Risk Profile; (b) Predicted risks curves