THE UNIVERSITY *of* EDINBURGH

Modelling and Predicting Medical Outcomes for Intensive Care Patients via Network Mechanisms and Machine Learning

Jorge Alejandro Gaete Villegas

Doctor of Philosophy

Artificial Intelligence Applications Institute
School of Informatics
University of Edinburgh

2024

Abstract

The intensive care unit (ICU) provides critical life support to a diverse group of patients with varying care needs, medical background, and demographics. This heterogeneity significantly increases the complexity of care, impacting the assessment of patients, the efficacy of treatments, and ultimately, the medical outcomes of these patients. To support medical decisions, ICU patients are constantly monitored, making available a wealth of patient information.

Recent work has extensively explored the application of machine learning to leverage this data to support decision-making in the ICU. However, challenges related to the quality of ICU data, patient heterogeneity, and the integration of medical and computer science knowledge create a gap between machine learning research and its practical implementation in critical care.

In this thesis, we introduce an innovative approach that combines a network representation of clinical data with non-parametric community detection in networks. With this method, we aim to complement and enhance existing models, addressing the aforementioned challenges related to ICU care.

We apply our approach to two tasks in the ICU: the identification of multimorbidity profiles for patients and in-hospital mortality prediction. Through these two studies we concretely present the challenges to the application of machine learning, show the limitations of existing approaches, and present the methodological benefits of our approach.

Our results highlight several advantages of our approach over existing methods. Firstly, the network representation of patients along with community detection, reveals statistically robust network structures that mitigate the impact of missing data. Secondly, our approach captures elements of heterogeneity, such as ethnicity and age, and automatically incorporates them into our clustering and prediction models. Third, the hierarchical structure of our approach offers flexibility to accommodate and adapt to varying levels of available data. Fourth, the non-parametric nature of our approach eliminates the need for complex parameter selection. Finally, the network representation of data offers a visual and more intuitive delivery of our models.

In this way, our work represents a step forward in the use of networks and machine learning methods to enhance decision-making support in the ICU in consideration of its complexity. Future directions outlined in this thesis involve enriching clinical data representation, exploring network reconstruction for mortality prediction, and incorporating medical knowledge to augment model interpretability.

Lay Summary

In the intensive care unit (ICU), where patients receive critical life support, providing effective care is challenging due to the diverse needs, medical backgrounds, and demographics of patients. To aid in decision-making, patient data is continuously monitored, offering a wealth of information. While machine learning has been explored to assist in ICU decision-making, challenges like data quality and patient diversity hinder practical implementation.

This thesis introduces a novel approach that combines network representation of clinical data with community detection to enhance existing models and address ICU care challenges. The method is applied to two ICU tasks: identifying patient multimorbidity profiles and predicting in-hospital mortality. Through these studies, the limitations of existing methods are highlighted, and the benefits of the proposed approach are demonstrated.

The results showcase several advantages: the robust network representation mitigates the impact of missing, captures heterogeneity factors like ethnicity and age on mortality, offers flexibility to adapt to varying availability of data, eliminates complex parameter selection, and provides a visually intuitive model delivery.

This work represents progress in leveraging networks and machine learning to support decision-making in the complex ICU environment. Future research directions include refining data representation, exploring network reconstruction for mortality prediction, and enhancing model interpretability with medical knowledge.

Acknowledgements

Although the endeavor of completing a doctorate can be lonely, it would be unbearable without the people who provide support during tough times. Here, I want to thank all those who helped me along the way and now share this moment of celebration.

I would like to start by thanking Jacques Fleuriot, my supervisor, and Petros Papapanagiotou, my co-supervisor at the beginning of the program. They began by accepting me into the program after just an interview call and reading a proposal. The faith they placed in me had the greatest impact on my life and that of my family. Thanks to their leap of faith, we moved halfway around the world, met incredible people, and discovered Edinburgh, a city that is now in our hearts.

The pandemic occurred midway through my PhD, and amongst the many changes it brought, it also brought Valerio Restocchi as my co-supervisor. Valerio not only introduced Network Sciences into my research but also brought a renewed energy that gave me the final push I needed to complete my PhD journey.

A PhD is also a time of personal growth and development, and I was fortunate to be surrounded by wonderful people with whom I shared my time in and around the Informatics Forum. First, I want to thank Jacques again, who was there to listen when I needed to be heard and to call me to my senses when necessary. To Valerio, for his patience and sincere interest in my development. To my initial lab-mates Imogen, Jake, and Mark, thank you for giving me a warm welcome to the lab. To my later lab-mates Zonglin, Fiona, and Nijesh, thank you for the unforgettable chats during hard-working hours. To Guillermo, Adarsh, and James, who helped me make sense of my thesis, and to Philip, Lauren, and Ricardo, who generously took the time to review the final draft, I extend my deepest appreciation. We not only shared research, but also many pints, lab meetings, seminars, parties, failures, and successes. I am extremely grateful to have shared this part of my life with all of you over these years, and I wish you nothing but the best in your future projects.

I cannot forget to thank my parents, Josefina and Jorge, for, almost annoyingly, always having faith in me. And I would like to remember Juana, Josefina, Mario, and Jorge. You might not be here, but you are always with me.

Finally and most importantly, I want to express my love and gratitude to Ksenia and Mila. My wife and daughter provided support and inspiration at every step. They endured my mood swings, my insecurities, and my talks about things that probably held little interest for them. In both, I found a connection to the important things in life, guiding me whenever I felt lost. Without you, this would not have been possible.

Ethics Statement

Ethical approval for the MIMIC-III project was granted by the Institutional Review Boards of Beth Israel Deaconess Medical Center (Boston, MA) and the Massachusetts Institute of Technology (Cambridge, MA). The requirement for individual patient consent was waived because the project did not impact clinical care, and all protected health information was de-identified.

(*Jorge Alejandro Gaete Villegas*)

Contents

Chapter 1

Introduction

The intensive care unit (ICU) provides life-critical support to a wide spectrum of patients, ranging from accident victims to individuals with chronic conditions, who span various age groups and can have multiple pre-existing conditions. This heterogeneity significantly increases the complexity of care, impacting the assessment of patients, the efficacy of treatments, and ultimately, the medical outcomes of patients (Forte et al., 2019).

To support decision-making in the ICU, patient conditions are continuously monitored. A wealth of data is generated as a result, offering opportunities to improve patient care. This abundance of data has driven research into applying machine learning to enhance specialized and timely attention in the ICU (Johnson et al., 2016b).

However, the quality of ICU data, heterogeneity of patients, and integration of medical and computer science knowledge create challenges for the effective use of machine learning in the ICU. Our work addresses these challenges by leveraging network-based data representation and models and focuses on the following research questions:

- **RQ1**: How can we improve the robustness of machine learning models over missing and sparse ICU data?
- **RQ2**: How can we enhance machine learning models to incorporate the heterogeneity of ICU patients?
- **RQ3**: How can we leverage available ICU data to simplify the training and validation of machine learning models?

In this chapter, we review the various challenges in more detail, their impact on the use of machine learning in the ICU, and their relevance to our research questions. We then present our approach and its benefits, followed by an outline of the rest of the thesis.

1.1 Challenges in applying machine learning in the ICU

Current literature extensively explores the use of machine learning for medical decision support in the ICU. Particularly, prediction and clustering methods have shown success in use cases such as patient risk assessment and multimorbidity profiling. However, the characteristics of ICU data, the diversity of patients, and the complexities of bridging machine learning models and medical requirements pose challenges for the use of machine learning. This section explores these challenges and their implications for the reliability and effective use of machine learning in the ICU.

1.1.1 Missing, sparse and erroneous data in the ICU.

Data captured in the ICU has the main goal of enhancing patient care rather than retrospective analysis (Johnson et al., 2016b). This poses a challenge as data might not be suitable or ideal for the application of machine learning algorithms (Ghassemi et al., 2020). Dealing with this issues is fundamental as it can significantly impact the quality and validity of machine learning models, analyses and findings.

A first source of concern is erroneous data, which refers to values that inaccurately reflect true measurements due to errors in measurement, recording, or processing . This is usual in the ICU, as readings can be affected by interventions. For example, elevated potassium levels may result of medical interventions rather than linked to the patient's actual health state (Johnson et al., 2016b).

Missing data is a significant concern often overlooked in ICU literature (Vesin et al., 2013). It occurs when intended data points are absent from datasets (Johnson et al., 2016b). This issue is typically categorized into three types: missing completely at random, where data absence is unrelated to other variables; missing at random, where data absence relates to observed variables; and missing not at random, where data is systematically missing. All types have the potential to diminish model performance (Sterne et al., 2009). In Chapter 5, we address this by presenting a mortality prediction model capable of handling missing data without requiring imputation.

The sparsity of data poses another significant challenge in the ICU, particularly for clustering methods. Sparse data refers to datasets with many zero or null entries, indicating infrequent occurrences of certain measurements or interactions within the dataset. Clinical data shows high dimensionality due to the wide range of possible conditions, but often sparse within individual patients due to the limited number of co-occurring conditions (Barnett et al., 2012b). As we discuss in Section 2.2, traditional

clustering algorithms struggle to effectively handle sparse data, potentially hindering the unveiling of clinically relevant patient subgroups.

Related work tackles data issues employing various strategies, mostly stemming from statistical tools in consideration of the specific case of missing data. For data missing at random, researchers rely on other features to draw relationships between the missing values and the available data. For example, patient clustering has been proposed to impute missing data based on the average value observed in other patients clustered together (Venugopalan et al., 2019). In cases where the data is missing not at random, and the missing data depends on other missing values, related work makes use of mixture models to uncover latent, non-observable, variables to rely on and conduct imputation (Michiels et al., 2002). While useful, these approaches bring an extra burden to the application of machine learning in the ICU.

To tackle these challenges, our work on **RQ1** seeks to develop machine learning methods capable of directly handling erroneous, missing, and sparse ICU data. By doing so, we aim to enhance the performance and reliability of machine learning models, contributing to improved decision-making in critical care.

1.1.2 Handling patient heterogeneity in the ICU

The ICU population is highly diverse due to their wide array of medical conditions, demographics, ongoing treatments, and admission causes. This heterogeneity can lead to varying medical outcomes, such as mortality, between seemingly similar patients receiving identical treatments (Forte et al., 2019). To tackle the heterogeneity, the identification of homogeneous groups of patients comes as a natural strategy. Recent literature has leveraged machine learning to identify clinically relevant groups of patients with acute respiratory distress syndrome (Reddy et al., 2020) and similar long-term disease profiles in the ICU (Zador et al., 2019).

However, the presence of heterogeneity poses challenges to the use of machine learning. It limits the available data for training machine learning models as samples are subdivided into more specific cohorts (Maslove et al., 2017). Furthermore, the heterogeneity of patients can conceal differences in outcome prevalence and data imbalance among patient cohorts in the whole population.

Related work addressing the need to account for heterogeneity often splits data into patient cohorts using clinical knowledge before analysis. Closely related to our work in the chapter 5, we can find how previous work aims to calibrate mortality predic-

tive models to specific populations. One example is the creation of a mortality index (SAPS, discussed in Section 2.1.4) where different indices were developed for populations of different ethnicities (e.g. Australasia, Central and South America, Eastern Europe). Another example can be found in the use of networks to describe the co-occurrence of clinical conditions for sex-specific cohorts (Kalgotra et al., 2017). While considering the heterogeneity of patients, this approach limits the data-driven discovery of new elements of homogeneity and effectively reduces the amount of available data for training models by splitting them into smaller samples.

To address the challenges related to patient heterogeneity, our work on **RQ2** aims to develop models capable of identifying and incorporating relevant dimensions of homogeneity amongst patients. By doing so, we seek to maximize the use of existing data while accounting for the diversity within the ICU population.

1.1.3 Integrating medical and machine learning knowledge

The use of machine learning to leverage large data collections to gain new insights into patients' conditions has been widely explored in literature. Clustering techniques such as K-meas and LCA (see Section 2.2) have been used to identify groups of patients with similar mental (Trevithick et al., 2015) or co-occurring conditions Zador et al. (2019). Other predictive models such as xg-boost have been used to model mortality for ICU patients (El-Manzalawy et al., 2021).

In these cases, machine learning is used to account for complex relationships between clinical variables and integrate medical knowledge to effectively support medical experts (Johnson et al., 2016b). However, these models do not always align directly with the characteristics of the ICU context. Therefore the use of machine learning requires specialized knowledge, which can be challenging in the clinical environment where access to machine learning expertise is limited (Callahan and Shah, 2017).

One of the challenges of applying machine learning in the ICU is aligning medical data with the statistical assumptions of the models. For example, methods such as agglomerative clustering (see Section 2.2) typically assume normal distributions of features, finding clusters with spherical shapes where data points are roughly equidistant from a central point. This is limiting because clinical data can form clusters with irregular shapes due to complex relationships between clinical features. These assumptions need to be considered to accurately reflect clinical reality Busija et al. (2019).

Another challenge involves hyperparameterization in various machine-learning ap-

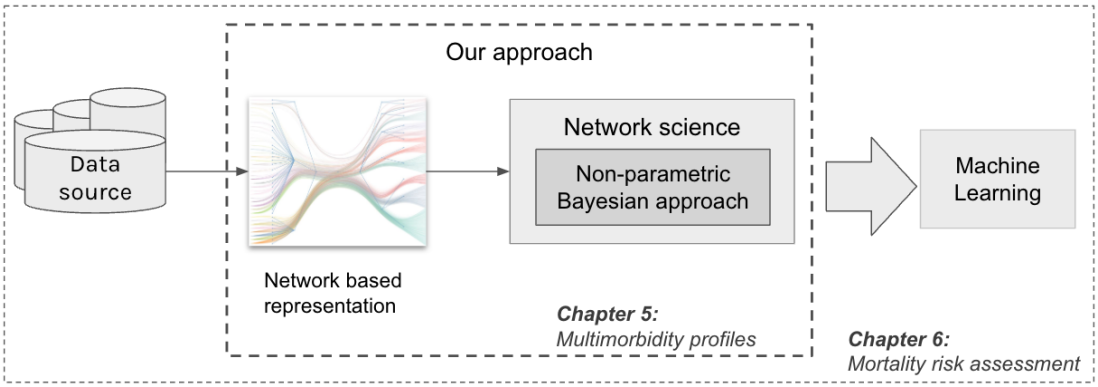

Table 1.1: Illustration of our approach in the context of this thesis. We include a mapping to the main chapters treating each element.

proaches. Widely used methods such as K-means clustering models require previous knowledge of the number of patient clusters present in the population. While heuristics exist to assist parameter selection, there is no consensus as to how to do model selection in a principled way (Raykov et al., 2016). Additionally, imposing the number of clusters can restrict the method's ability to discover novel and unexpected insights.

Our work on **RQ3** aims to tackle the challenges stemming from the technicalities involved in the application of machine learning within the ICU. By leveraging non-parametric statistical models we expect to lift some of the complexities involved in the training and validation of machine learning. By doing so, our goal is to facilitate the practical use of machine learning models in the ICU context.

1.2 Our approach

In our work, we combine network science and non-parametric techniques to explore answers to our research questions and address ICU-specific challenges. Our approach is composed at its core of three pillars: Network-based representation of patient data, network science to unveil robust structures within this network and a non-parametric Bayesian approach for a data-driven definition of our models. In this section, we briefly introduce each of these elements.

1.2.1 Network-based modelling

The use of tabular data to represent patients and their clinical features is common for current clustering and prediction models used in the ICU (see Sections 2.2 and 2.4).

While this representation aligns with the structure and format of medical records, it can struggle to naturally describe the complexities of the relationships between patients and their features (Doreian et al., 2019). In contrast, the use of network representations of medical data offers the advantage of directly encoding complex dependency relationships between features e.g. it can encode the impact of cellular components on other cells across organs (Barabási et al., 2011). We believe that this benefit can extend to the ICU, where the heterogeneity of patients leads to intricate relationships between patients. Additionally, the graphical nature of networks provides a more intuitive representation for understanding the data (Vellido, 2020a).

1.2.2 Use of network science

Network science aims to discover structures, patterns and behaviours within complex networks with a focus on the relationships between its components (Newman, 2018). A first benefit of network science models is their resilience to missing data, since the absence of links between some nodes does not necessarily disrupt the overall connectivity patterns observed in the network (Barabási, 2013). The focus on relationships allows network science models for community detection to detect cluster structures even when data is sparse and when communities are relatively small compared to the overall population (Peixoto, 2017). Furthermore, these models offer a natural visual representation in networks. As discussed in Chapters 4 and 5, this allows for graphical visualisation of the inferred models, enhancing their intuitive presentation (Vellido, 2020a)

1.2.3 Use of non-parametric methods.

A common element across machine learning methods is the need for complex model training and selection steps. This is expressed in various ways, for instance, the need to determine the expected number of groups in clustering algorithms (see Sections 2.2) or the optimal number of trees in a predictive ensemble method (see Section 2.4). From a computer science standpoint, these decisions are not straightforward and involve significant technical work. However, driving these decisions from a purely medical perspective can limit the potential contributions of data-driven insights to medical knowledge. The use of non-parametric methods to analyse network structures, discussed in Section 2.3.3, allows us to automate the selection of models with a data-driven and statistically principled approach. Recent literature has emphasized the advantages

of non-parametric Bayesian approaches in addressing the intricacies of modelling in healthcare contexts (Johnson et al., 2016b).

1.3 Opportunities for machine learning in the ICU

Existing literature discusses many promising opportunities for the effective use of machine learning, such as the diagnosis and treatment of patients (Ghassemi et al., 2020), the integration of multiple medical data sources (Johnson et al., 2016b), and the advancements in personalized medicine (Hu et al., 2016). However, the challenges of the ICU discussed in Section 1.1, limit the use of machine learning within that setting.

In this section, we introduce two relevant tasks in the ICU, which will serve as concrete examples to highlight the limitations of current approaches and to demonstrate the effectiveness of our approaches in answering our research questions. These tasks will be covered in more detail in Chapters 4 and 5.

	Multimorbidity profiling	Mortality prediction
Nature of the clinical data	• Sparse data	• Missing data.
Patient heterogeneity	• Complex features interaction in different cohorts.	• Different prevalence per cohort. • Unbalanced data.
Medical and ML knowledge integration	• Complex model selection. • Unknown number of clusters.	• Complex model training. • Reduced samples size.

Table 1.2: ICU Challenges to the application of machine learning and their implications on relevant ICU tasks.

1.3.1 Multimorbidity profiling in the ICU

With an ageing global population, the ICU faces an increasing number of patients with one or more co-occurring long-term health conditions, known as multimorbidity (Barnett et al., 2012a). This growing phenomenon means additional risk factors for patients and challenges to care providers due to increased diversity in the ICU-admitted population. Consequently, the identification of patient groups with similar conditions, known as multimorbidity profiles, has become increasingly important (Busija et al.,

2019). Detecting these multimorbidity profiles is crucial for tailoring the treatment of patients and increasing the effectiveness of care, optimizing resource allocation, and improving risk assessment (Prados-Torres et al., 2014).

However, applying machine learning to this task comes with various challenges. Firstly, is the intrinsic sparsity of morbidity data. While the number of possible morbidities can reach up to 30 (Elixhauser et al., 1998), depending on the definitions used, each patient typically presents only a limited number (Barnett et al., 2012a). This sparsity poses challenges for current clustering methods in finding meaningful and/or unbiased clusters. Secondly, although medical literature indicates that both the types and number of conditions influence a patient's condition, existing work often fails to consider these factors simultaneously. While more morbidities generally increase mortality risk (Busija et al., 2019), some work finds single morbidities more impactful than multiple combined ones. For example, the Elixhauser Index shows that a solid tumor alone has a greater impact on mortality risk than the combined presence of alcohol abuse and liver disease (Elixhauser et al., 1998). Lastly, model training process involves complexities, particularly in determining parameters such as the optimal number of profiles. As discussed in Section 2.2, current methods rely on heuristics which may not consistently yield meaningful or replicable results (Raykov et al., 2016).

In Chapter 4 we explore multimorbidity profiles in a study cohort extracted from the MIMIC-III dataset, itself presented in Chapter 3. We identify these profiles via common methods in literature and contrast them to the results obtained using our approach. By doing this we explore how our community detection approach yields clustering results that successfully tackle the challenges presented above.

1.3.2 Mortality prediction in the ICU

Predicting mortality in the ICU is another critical task where machine learning presents great potential (Johnson et al., 2016b; Ghassemi et al., 2020). Accurate mortality prediction enables early intervention and also plays a pivotal role in resource allocation, optimizing patient care, and enhancing risk assessment (Vincent and Moreno, 2010).

However, applying machine learning to this task presents its own challenges. Firstly, imbalances in medical outcomes, such as mortality, can result in biased prediction models. Secondly, the diversity among ICU patients poses a challenge to create machine-learning models able to account for all cohorts of patients. For example, underrepresented cohorts have a smaller sample size in the population, affecting the training

of predictive models on these groups. Additionally, the interpretability of machine learning models is crucial in clinical decision-making, and striking the right balance between model complexity and interpretability is a challenge.

In Chapter 5, we look into mortality prediction in the context of the ICU, leveraging validated medical knowledge, community detection techniques and machine learning to develop a hierarchical mortality prediction models. By doing so, we address the challenges posed by ICU data using an adaptable approach to predicting mortality for ICU patients.

1.4 Thesis outline

In the next chapters, we extensively examine the application of our approach in the ICU context and its effectiveness in answering our research questions. In Chapter 2, we introduce various healthcare concepts that are both recurrent in our work and fundamental to understanding the context of the dataset and results. In the same chapter, we introduce relevant clustering algorithms in the healthcare domain and contrast them to the use of community detection in networks. In particular, we examine the hierarchical stochastic block modelling, as an alternative to traditional clustering. Chapter 3 introduces the MIMIC-III dataset, the database used throughout this research, and presents the process followed to extract our study cohorts.

In Chapters 4 and 5 we detail our approach to addressing multimorbidity profiling and mortality risk prediction, respectively. In Chapter 4, we present a novel application of community detection for multimorbidity profiling in the ICU. This approach offers significant advantages such as the simplification of the model construction via the automatic selections of the model parameters, the incorporation of a complex mix of dense and sparse features in multimorbidity profiles, and a layered profile structure that provides a more detailed description of multimorbidity groups. In Chapter 5, we build on the advantages of hSBM to transform SAPS-II features, a mortality risk score, into a simplified input space which we leverage to build a hierarchical mortality prediction model. This system presents advantages such as the lack of patient information imputation, adaptability to varying availability of data, and a more stable behaviour across a different cohort of patients. In Chapter 6, we summarize our contributions, examine some limitations and post future research avenues.

Chapter 2

Background

The Intensive Care Unit is a challenging environment combining a situation with critically-ill patients requiring immediate life-saving care with the need to allocate limited human and technical resources. In this environment, patients are continuously monitored, generating a wealth of data, which constitutes a valuable source of information that can assist medical staff in assessing the condition of patients, deciding on resource allocation, supporting research efforts and evaluating the ICU performance.

Machine learning has emerged as an important tool to enhance the ability of healthcare professionals to provide care. Different classification and prediction methods have been proposed to support different tasks in the ICU such as patient stratification and prognosis. However, a disparity exists between machine learning models developed for ICU predictions and the available ICU data, which primarily serves the purposes of patient treatment and administrative management rather than machine learning research.

In our work, we aim to bridge this gap by making use of various elements from the machine learning and healthcare domains. In this chapter, we introduce these concepts, which we later apply in the construction of models that integrate healthcare knowledge while addressing the challenges of the ICU setting. We start by presenting healthcare concepts relevant to our work and their derivation in current healthcare datasets. Then we introduce some of the machine learning algorithms commonly used in healthcare research. We follow this with the introduction of community detection in networks as complimentary to the traditional machine learning approaches. Finally, we discuss widely used metrics used for model evaluation in the healthcare domain.

2.1 Healthcare concepts relevant to our research

In this section, we introduce concepts from the healthcare domain that are relevant to our work. We start by presenting the International Classification of Diseases, a widely used convention to describe patients conditions. Based on this classification we introduce the Charlson and Elixhauser classification of patient´s long-term conditions, known as comorbidities. Finally, we introduce Severity of Illness Scores, which are numeric systems designed to asses a patient's condition based on specific clinical variables such as temperature or heat rate.

2.1.1 International Classification of Diseases.

A first challenge in patient monitoring is agreeing on medical terminology, including diagnoses and treatments. The International Classification of Diseases (ICD) is a standardized coding system aimed at tackling this issue. It is used to classify and describe medical conditions, diseases, injuries and other health-related issues such as interventions and medication (NCHS, 1980).

We focus on the ICD-9 version in our research, despite the latest version being ICD-10. This is because the Medical Information Mart for Intensive Care III dataset (MIMIC-III), fundamental for our work and presented in Chapter 3), employs this version. The ICD-9 utilizes up to 6 digits to describe a patient diagnosis (e.g. ICD 140: Malignant neoplasm of the lip) including details of a wide variety of signs, symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or disease (e.g. ICD 140.6: Malignant neoplasm of the commissure of the lip).

It is worth noticing that while advantageous for medical research (Cui et al., 2018; Zador et al., 2019), the ICD is designed for administrative purposes mainly. Nonetheless, it remains a valuable and widely utilized tool for algorithmically computing various healthcare-related concepts (Johnson et al., 2018), such as sepsis and comorbidities presented in the next section.

2.1.2 Charlson and Elixhauser Comorbidities indices

Patients admitted to hospitals can suffer chronic conditions prior to their presentation (Johnson et al., 2018). Such conditions, known as comorbidities, are disorders unrelated to the primary reason for hospitalization that can have a significant impact on the precision of patient assessment, the efficacy of treatment, and medical outcomes

Comorbidity	score	Comorbidity	score	Comorbidity	score
aids	0	Diabetes uncomplicated	0	Paralysis	4
Alcohol abuse	0	Drug abuse	-11	Peptic ulcer	0
Blood loss anemia	-3	Fluid electrolyte	11	Peripheral vascular	4
Cardiac arrhythmias	8	Hypertension	-2	Psychoses	-6
Congestive heart failure	9	Hypothyroidism	0	Pulmonary circulation	5
Chronic pulmonary	3	Liver disease	7	Renal failure	7
Coagulopathy	12	Lymphoma	8	Rheumatoid arthritis	0
Deficiency anemias	0	Metastatic cancer	17	Solid tumor	10
Depression	-5	Other neurological	5	Valvular disease	0
Diabetes complicated	1	Obesity	-5	Weight loss	10

Table 2.1: Elixhauser comorbidity index and score. All comorbidities are listed along with their score indicating their impact on a patient's mortality probability.

(Charlson et al., 1987; Elixhauser et al., 1998). Multimorbidity, or the presence of multiple comorbidities (Barnett et al., 2012b), is a significant factor contributing not only to a higher likelihood of adverse medical outcomes but to an increased resource utilization (Forte et al., 2019; Reddy et al., 2020; Busija et al., 2019).

These indices, relying mostly on medical consensus, group patient diagnoses (as ICD-9 codes) into clinically relevant groups, known as comorbidity indices. Since these indices are defined using the ICD coding system, they can be calculated with administrative data and creating a standard that facilitates their study across different electronic health record. The Elixhauser index lists 31 comorbidities, while the Charlson index comprises 18. Beyond offering a detailed account of comorbidities, both indices provide a systematic approach to calculate a patient's mortality risk. This involves summing up the scores corresponding to the contribution of each comorbidity to provide an overall risk mortality scores. Tables 2.1 and 2.2 present the Elixhauser and Charlson scores respectively.

We take advantage of these indices due to their widespread use and relevance in healthcare research (Austin et al., 2015; Johnson et al., 2018; Charlson et al., 2022). We will employ them to identify multimorbidity profiles using traditional machine learning methods as described in Section 4.2.2 and community detection as covered in Section 4.2.4. Additionally, we provide a detailed description of these indices in our study cohort in Chapter 3.

Comorbidity	Score	Comorbidity	Score
* Mild liver disease	1	Aids	6
* Severe liver disease	3	Congestive heart failure	1
* Diabetes with complications	2	Peripheral vascular disease	1
* Diabetes without complications	1	Cerebrovascular disease	1
* Malignant cancer	2	Dementia	1
* Metastatic solid tumor	6	Chronic pulmonary disease	1
Paraplegia	2	Rheumatic disease	1
Renal disease	2	Peptic ulcer disease	1
* Only highest score considered for score		Myocardial infarct	1

Table 2.2: Charlson comorbidity index and score. All comorbidities are listed along with their score indicating their impact on a patient's mortality probability.

2.1.3 Sepsis and its computation from medical records

Sepsis is defined as a life-threatening organ dysfunction caused by a dysregulated host response to infection (Johnson et al., 2018). This is a major and costly condition, linked to an increase in the mortality rate of patients in hospital (Angus et al., 2001). The study of sepsis is an important area of research, as predicting its onset and evaluating the effects of treatments can be of benefit to a patient's survival (Reddy et al., 2020; Seymour et al., 2019).

Of particular interest is the impact of sepsis on mortality, leading to the construction of tools to detect its onset on patients (Vincent et al., 1996b; Davies and Hagen, 1997) and to assess its impact on patients with multiple long-term conditions (Zador et al., 2019; Popoola et al., 2021). This makes sepsis a relevant element to our study of mortality in Chapter 4.

Unfortunately, the onset of sepsis is not typically documented in clinical records as it is difficult to capture (Johnson et al., 2018; Angus et al., 2001). One approach to this is to compute it retrospectively using the medical records of patients at discharge. A common algorithm to identify sepsis using ICD-9 coded information follows 3 criteria (Angus et al., 2001): explicit coding, the concurrent presence of infection (bacterial/fungal) and organ dysfunction, or if the patient shows infection (bacterial/fungal) and is under mechanical ventilation. This criterion has been validated (Iwashyna et al., 2014) and algorithmically implemented in the MIMIC-III database (Johnson et al., 2016d).

The Angus definition of sepsis represents a reasonable attempt to reconstruct the onset of sepsis from administrative data (Iwashyna et al., 2014). However, it has limitations. Coding accuracy and inconsistent practices may lead to incorrect assessments of sepsis onset (Johnson et al., 2016a). A significant limitation is the lack of physiological variables and the broad sepsis definition in the Angus criteria, which can lead to misclassification. For instance, organ dysfunction might not be related to infection, thus missing the sepsis onset (Cohen et al., 2015). Overall, methods using administrative data to reconstruct sepsis are effective for identifying healthy patients but have moderate sensitivity (Cohen et al., 2015).

2.1.4 Severity of illness scores

Severity of illness scores are systems designed to quantify the severity of a patient's condition. They are actively used in the prediction of medical outcomes, especially in intensive care (Vincent et al., 1996a; Johnson et al., 2013a). These scores map a series of clinical variables into a single numerical value to reflect the overall condition of a patient (El-Manzalawy et al., 2021). Besides assisting physicians, they are used for risk stratification in medical research, such as patient groups comparison in clinical trials (Johnson et al., 2018).

Their simplicity and transparency make these scores widely used in ICU settings (Zador et al., 2019; Johnson et al., 2013b). We thus use these as benchmarks for our work on mortality prediction in Chapter 5. In the current chapter, we focus on five widely used scores both in research and clinical settings: SOFA, SIRS, APS-III, OASIS and SAPS-II (El-Manzalawy et al., 2021; Strand et al., 2009; Kong et al., 2020; Moreno and Apolone, 1997).

The development of these scoring systems involves a series of steps. Firstly, a number of clinical variables are selected to use as predictors for the patient's condition. Researchers combine medical expertise, statistical analysis and machine learning to achieve an optimal selection of these variables. In cases such as SOFA (Vincent et al., 1996b), SAPS-II (Le Gall et al., 1993) and APS-III (Zimmerman et al., 2006) this process is mostly driven by medical consensus, while in others, such as for OASIS (Johnson et al., 2013a), it is assisted by machine learning in the form of a genetic algorithm that is used to select features from a pool of candidates. An essential characteristic of this stage is the reliance on medical knowledge to define clinically meaningful variables.

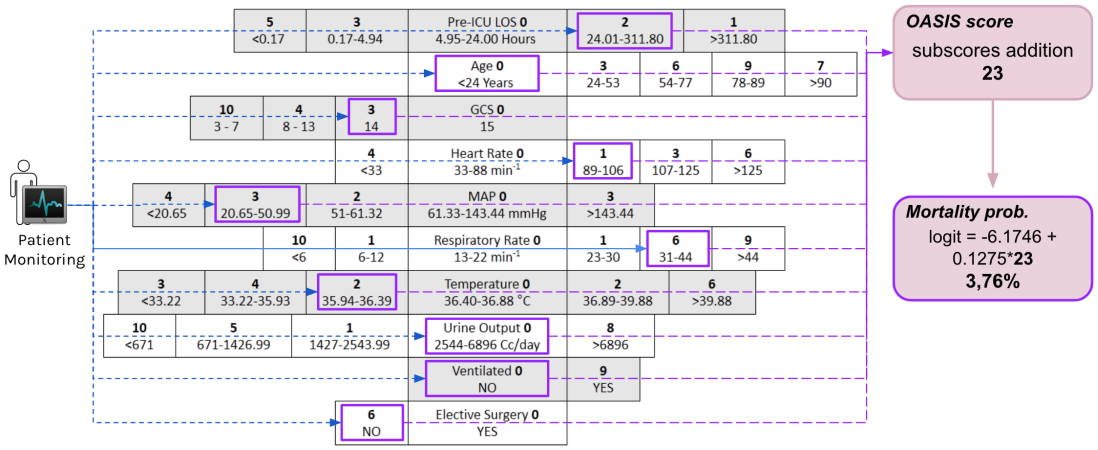

Figure 2.1: Computing the OASIS score and mortality risk. The patient's clinical variables are discretized and assigned a subscore according to the OASIS scoring system. These subscores are summed up to provide a severity score, of 23 in this example, and a mortality probability based on this score, of 0.037 in this example.

In the second phase, the clinical variables are further categorized into specific ranges, and corresponding weights are assigned to these ranges. The approach to achieving this can vary widely between scoring systems. In some cases, this process is purely based on medical knowledge, like for SOFA. In contrast, approaches such as SAPS-II or APS-III employ statistical analysis, using multiple logistic regressions and testing various models with different ranges to determine optimal weights. Alternatively, *particle swarm optimization* is used in OASIS to test randomly assigned weights to optimize the discriminatory power of the resulting model. Finally, models such as SAPS-II, APS-III and OASIS provide a transformation function from severity score to mortality probability.

Figure 2.1, we use OASIS to illustrate the usage of severity of illness scores, a system that we will revisit in Chapter 5. In this figure, we show how a patient's clinical variables are mapped to their corresponding range and assign a subscore as defined in OASIS (Johnson et al., 2013a). The OASIS score is calculated by adding all the patient's subscores, giving a score of 23 in the example provided in Figure 2.1. This OASIS score can be further processed to obtain an in-hospital mortality probability, which is 0.037 in our example. Among the scores considered in our work, only OASIS, SAPS-II and APS-III provide a function to transform scores into mortality probability. Below we introduce the severity scores that are further mentioned in Chapter 5. A summary of their clinical variable is presented in Table 2.3.

OASIS (Oxford Acute Severity of Illness Score) (Johnson et al., 2013a): Pre-

sented in Figure 2.1, is used to assess the severity of illness in patients over 18 years old. It incorporates, within the first 24 hours of admission, patients information on heart rate, mean arterial pressure, temperature, respiratory rate, urine output, pre-ICU admission length of stay, neurological state (using the Glasgow Coma Score), and age, plus the use of mechanical ventilator and admission type (elective/emergency). A mortality probability as a function of the patient's OASIS score is provided. The OASIS scoring system presented in Figure 2.1.

SOFA (Sequential Organ Failure Assessment) (Vincent et al., 1996b): Used to assess the severity of organ dysfunction in critically ill patients. It evaluates the functioning of six organ systems: respiratory, coagulation, liver, renal, cardiovascular and central nervous. Each system is scored on a scale from 0 to 4 based on the degree of dysfunction using the information collected within the first 24 hours of admission. We present the SOFA scoring system in Table A.1 in Appendix A.1.

SIRS (Systemic Inflammatory Response Syndrome) (Davies and Hagen, 1997): Used to identify systemic inflammation in response to various conditions, such as infection, trauma, or other inflammatory processes. It evaluates four criteria within the first 24 hours of admission: abnormal body temperature (either fever or hypothermia), abnormal heart rate, abnormal respiratory rate, and abnormal white blood cell count. The SIRS scoring system is presented in Table A.2 in Appendix A.1

SAPS-II (Simplified Acute Physiology Score II) (Le Gall et al., 1993): SAPS-II is used to predict the mortality risk of critically ill patients within the first 24 hours of their admission to an intensive care unit (ICU). It considers various physiological and clinical parameters, such as age, chronic health conditions, and laboratory values, to provide a mortality risk estimate. SAPS-II is valuable for guiding treatment decisions and assessing ICU performance. The transformation from SAPS-II to a mortality probability is provided as part of the scoring system. The SAPS-II clinical variables and subscores is presented in Table 5.1 in Chapter 5.

APS-III (Acute Physiology Score III) (Zimmerman et al., 2006): This scoring system is part of the APACHE-II system, and it is used to assess the severity of illness in critically ill patients. It evaluates twelve physiological parameters, such as heart rate, blood pressure, temperature, and laboratory values, to provide a numerical score that reflects the patient's overall condition. A mortality probability function is provided. the APS-III system is presented Table A.3 in Appendix A.1

Clinical Variable	SAPS-II	SOFA	SIRS	APS-III	OASIS
Demographic Variables					
Age (years old)	✓				✓
Vital Signs					
Heart rate (beats/min)	✓		✓	✓	✓
Body Temperature (°C)	✓		✓	✓	✓
Respiratory rate (breaths/min)			✓	✓	✓
Mean BP (mmHg)				✓	✓
Systolic BP (mm Hg)	✓				
Respiratory Function					
PaO2/FiO2 (mm Hg)	✓	✓			
A-a gradient				✓	
Patient is ventilated					✓
Renal Function					
Urine output (L/d)	✓				
Urine output (ml/24 hours)				✓	
Urine output (cc/24 hours)					✓
Serum Urea Level (L/d)	✓				
Creatinine (mg/dL)		✓		✓	
BUN (mg/dl)				✓	
Hematological Parameters					
White Blood Cell count (WBC) (10^3/cu mm)	✓		✓		
White Blood Cell count (WBC) (10^9/L)				✓	
Platelets (10^3/mm^3)		✓			
Hematocrit (%)				✓	
Electrolyte and Metabolic Balance					
Serum potassium (mmol/d)	✓				
Serum sodium level (mmol/L)	✓			✓	
Serum bicarbonate level (mWq/L)	✓				
Glucose (mg/dl)				✓	
Acid-base				✓	
Liver Function					
Bilirubin level (micro mol/L)	✓	✓			
Bilirubin (mg/dl)		✓		✓	
Albumin (g/dl)				✓	
Neurological Function					
Glasgow Coma Score (GCS)	✓	✓		✓	✓
Other Variables					
Chronic diseases	✓				
Type of admission	✓				
Hypertension		✓			
Elective surgery					✓
Pre-ICU (hours)					✓

Table 2.3: Clinical variables assessed in the SAPS-II, SOFA, SIRS, APS-III, and OASIS severity scores, along with the percentage of missing data for each variable. Variables are grouped according to the assessment goal and marked with a check under the respective severity score if they are included in that score's assessment criteria.

2.2 Clustering methods.

Identifying homogeneous groups of patients is critical for dealing with the diversity of ICU patients. Clustering is the unsupervised process of discovering groups of similar objects (Jain et al., 1999), such that objects within each cluster share more similarities with each other than with objects in other clusters (Huang, 1997). By modelling patients as objects with clinical characteristics, clustering becomes a natural approach to leveraging the abundance of available clinical data to handle patient heterogeneity.

In this section, we briefly describe algorithms widely used in the context of multimorbidity profiling, a topic explored in Chapter 4. Additionally, we discuss model selection, a challenging aspect of the use of these methods. We aim to provide an understanding of these models and shed light on their limitations within the context of the ICU challenges presented in Chapter 1.

2.2.1 Partition-based clustering

We begin our review with partition-based algorithms using observed features to group similar data. These methods define "similarity" as the "distance" between observations in the feature space (Murphy, 2012). The goal is to find a partition that minimizes a clustering error, defined in terms of the distance between observations within clusters (Jain et al., 1999). As partition clustering is an NP-hard optimization problem (Ezugwu et al., 2022), multiple approaches exist to approximate a solution. To illustrate these algorithms, we focus on centroid-based partition algorithms due to their frequent use in the healthcare domain.

Consider a general scenario with N objects $X = X_1, \ldots, X_n$, with X_i as a vector of features like a patient's age or gender. Cluster-based algorithms aim to partition these N objects into K clusters, where K is fixed based on domain knowledge or statistical criteria. This is achieved by finding the assignment of X_i to clusters that minimize the overall distance between each point to the centre of its cluster, known as *centroid* (Huang, 1998).

$$E = \sum_{k=1}^{K} \sum_{i=1}^{N} y_{i,k} d(X_i, \mu_k) \tag{2.1}$$

Equation 2.1 shows a generalized form of this minimization problem, with $d()$ a distance function, μ_k the centroid of cluster k, and $y_{i,k}$ representing the cluster membership (1 if object X_i is in cluster k, 0 otherwise). This equation is minimized iteratively:

first, fixing μ_k and choosing $y_{i,k}$ to minimize E, and then fixing $y_{i,k}$ and recalculating μ_k. This process continues until a predefined threshold for E is met (Huang, 1997, 1998). K-means is one of the most popular centroid-based algorithms. Here, $d()$ is defined as $\|X_i - \mu_k\|^2$, the squared distance between each observation and its assigned cluster k (Jain et al., 1999). However, this method applies to numerical data, where "distance" is meaningful. K-modes (Huang, 1997), a modification of k-means, provides a solution for categorical data, also common in healthcare. In this case, $d()$ represents the number of matching variables between observations, and μ_k corresponds to the mode for each variable over the observations in each cluster (Huang, 1997, 1998).

These methods are popular for their simplicity and scalability but come with several limitations. First, the number of clusters must be known in advance, which is rarely the case (Rindskopf and Rindskopf, 1986). Second, clustering relies on the definition of the distance $d()$. While this can work for compact and uniform clusters (Jain et al., 1999), it will struggle to accommodate more complex clusters of patients with non-uniform feature distributions (Jain et al., 1999). Thirdly, the clustering model depends on the initialization of centroids, posing a challenge to the reproducibility of results. Finally, they assume independently distributed variables and complete and accurate data, with violation of this assumption leading to inaccurate and unexpected results.

2.2.2 Probabilistic partition-based methods

An alternative to distance-based methods is the use of probabilistic models to identify clusters based on the distribution of features rather than on their observed values. One such method is the Latent Class Analysis (LCA), on which we focus due to its significance in the healthcare domain (Forte et al., 2019; Busija et al., 2019; Hagenaars and McCutcheon, 2002). LCA is a finite mixture model, assuming that the observations are representations of an underlying structure of independent latent classes.

Let us extend the previous scenario to consider the distribution of the observed variables. Suppose that each X_i contains J categorical variables, each with R_j possible values. We define X_{ijr} as the observed value of feature j in object i, such that X_{ijr} = 1 if X_i presents the value r for feature j and 0 otherwise. LCA approximates the joint distribution of the J variables as a weighted sum of K latent classes, where K is a hyperparameter of the model like in the centroid models. We represent π_{jrk} as the probability that an object in class k exhibits the value r for feature j, therefore within a class $\sum_{r=1}^{R_j} \pi_{jrk} = 1$. The probability distribution of an object X_i in class k having a

particular set of values r, assuming independence given its class membership, is given by (Linzer and Lewis, 2011):

$$f(X_i|\pi_k) = \Pi_{j=1}^{J}\Pi_{r=1}^{R}(\pi_{jrk})^{X_{ijr}} \tag{2.2}$$

Finally, the membership probability function for an observation is derived by incorporating p_k, the probability of an object being part of class k:

$$P(X_i|\pi, p) = \sum_{k=1}^{k} p_k f(X_i|\pi_k) \tag{2.3}$$

The model parameters to estimate are the probabilities p_k and π_{jrk}. To fit the final model, a common strategy is to maximize the log-likelihood of equation 2.3 for all observations X_i. Algorithms like the expectation-maximization are often employed for this task. It is worth noticing that the number of independent variables to estimate grows rapidly with J, R, and K (Linzer and Lewis, 2011).

By leveraging a probabilistic model, LCA is a more robust approach to missing and erroneous data than centroid-based methods. However, its use still presents drawbacks. Firstly, it lacks a principled method to determine the optimal number of clusters, similar to simple methods such as K-modes. Secondly, as the dataset grows in complexity so does the number of parameters to estimate, bringing scalability and usability challenges. Finally, the class independence assumptions in mixture models (see equation 2.2) and LCA, in particular, might not hold in the healthcare context, and might result in unexpected behaviours (Busija et al., 2019).

2.2.3 Hierarchical partition-based clustering

The methods previously discussed produce flat clustering models. In contrast, hierarchical methods create a nested hierarchy of clusters and are often used in healthcare for their enhanced view of the data (Busija et al., 2019). Two are the main approaches to hierarchical clustering (Murphy, 2012): **Agglomerative** and **divisive**. We focus on the former to illustrate this approach due to its relevance in the medical domain.

The agglomerative method groups clusters based on their similarity, measured as the distance, e.g. Euclidean, between their members. The process begins with each data point as its own cluster and then iteratively merges the closest clusters until only one cluster remains (Murphy, 2012). The calculation of the distance between clusters leads to three variants of agglomerative clustering: **Single link**, the distance between

two clusters is the distance between the two closest observations in each cluster; **Complete link**, the distance between two clusters is the distance between the two furthest observations in each cluster; and **Average link**, the distance between two clusters is the average distance between all the observations in each cluster.

The process can be represented as a tree, called a dendrogram, with observations on the horizontal axis and the metric of closeness on the vertical axis (Murphy, 2012). The initial N clusters are at the bottom of the tree, and as they are merged they are joined in the tree. The height of a branch represents the dissimilarity between the groups merged at its leaves. Dendrograms are often used for model selection, as discussed in Section 2.2.4.

Amongst the benefits of this approach are the visual representation of the clustering structure and its deterministic nature, with no dependency on the initialization of the algorithm. However, its use comes with significant drawbacks. First, as with the previous methods, there is no principled way to determine the number of clusters. Second, as in centroid-based methods, distance metrics may not work well with large feature input spaces with many values being zero or missing. Third, similarly to other partition methods (see Section 2.2.2, the quality of clusters relies heavily on the observed data which can be of low quality. Last, the computation of distances becomes increasingly expensive as the number of observations and features increases.

2.2.4 Model selection

The lack of a principled way to define the optimal number of clusters in the approaches discussed makes selecting the best-fitting model challenging. In this section, we present various mechanisms used in the process of model selection, to choose one partition when there are multiple candidates.

In agglomerative clustering, model selection is typically done by maximizing the dissimilarities between clusters. A common heuristic is to use a dendrogram (see Section 2.2.3) and make a cut at the point where a visible gap in the height of the branches is observed. This gap shows the maximum dissimilarity between clusters. Unfortunately, this gap is not always visible or clear (Murphy, 2012). We come back to this problem at the end of this section.

In the case of partition clustering, a common approach is to generate candidate models for a predefined set of clusters (e.g. 5, 6, and 7 clusters) and then compare them using a quality criterion. The optimal number of clusters is determined by choosing

the model that minimizes the criterion. For partition methods, such k-modes (Section 2.2.1), the quality is typically measured as proportional to the distance between each element in a cluster and the cluster's centroid, known as inertia or distortion (Huang, 1997; Rousseeuw, 1987). In Chapter 4 we use the distortion as a metric to evaluate the quality of k-modes and agglomerative clustering models. This represents the average distance of each observation to the centroid of the cluster it has been assigned to. Equation 2.4 presents the distortion of a model assigning a total of N observations to K possible clusters.

$$Distortion = \frac{1}{N} \cdot \sum_{n=1}^{N} \min_{j \in K} (x_i - \mu_j)^2 \tag{2.4}$$

In the case of LCA (Section 2.2.2), the error is measured in terms of the likelihood of the clustering model given the observed data (Murphy, 2012). Two widely used likelihood criteria are the Bayesian Information Criteria (BIC) (Schwarz, 1978) and the Akaike Information Criteria (AIC) (Akaike, 1998), shown below.

$$\begin{aligned} BIC &= -2 \cdot \ln(\hat{L}) + K \cdot \ln(N) \\ AIC &= -2 \cdot \ln(\hat{L}) + 2K \end{aligned} \tag{2.5}$$

Were $\hat{L}$ denotes the maximum likelihood estimate of the model, k the number of parameters in the model, and N the number of observations. Both criteria score models in a similar fashion, by rewarding the likelihood of a given model while penalizing for its complexity, e.g. the number of clusters, as shown in equation 2.5. It has been reported that AIC performs better with smaller samples while BIC can be more useful with relatively larger samples (Burnham and Anderson, 2004). As a result, both BIC and AIC are usually presented together as evidence for model selection.

Relying solely on error metrics as a selection criterion continues to pose challenges, for example, multiple partitions with different numbers of clusters might yield similar metrics (Åkerlund et al., 2022; Burnham and Anderson, 2004). A common heuristic to decide the optimal number of clusters utilizing the error criteria is to identify the "elbow point" in a criterion versus the number of clusters curve. The elbow point is that point from which adding a new cluster does not decrease the quality criteria significantly but could lead to overfitting, for example having very small clusters (Tibshirani et al., 2001). A formal approach to identify the elbow is presented by Satopaa et al. (Satopaa et al., 2011), where the elbow is defined as the point of maximum curvatures in the error versus number of clusters curve.

The elbow method is a useful tool to achieve a more principled model selection among a fixed set of models. Particularly in cases where the candidate models do not depend on initialization conditions, such as in agglomerative clustering. However, it leaves us with one extra concern when the models depend on the initial condition such is the case for K-modes and LCA. In this case, multiple solutions exist for each number of clusters, and so the question of selecting the optimal solution for each cluster number persists. To address this challenge, in Section 4.2.2 we describe multiple experiments. We apply the elbow method to determine the best solution from a diverse set of possible methods and then decide the optimal number of clusters based on a frequency analysis.

2.3 Community detection for clustering

As described, traditional methods like K-modes and hierarchical clustering use tabular data representations to calculate differences between patients' characteristics (Sections 2.2.1 and 2.2.3). LCA uses vector representations to assign patients with membership probabilities to clusters (Section 2.2.2). In contrast, community detection relies on a network representation of data to uncover structural components like groups of similar nodes. Instead of relying on pairwise similarities, community detection defines node similarity based on their connections within the network (Javed et al., 2018). This approach distinguishes community detection from traditional clustering methods, and significant efforts have explored its suitability in healthcare (Barabási et al., 2011).

In this section, we introduce two key aspects of our work with community detection. First, we discuss the network representation used throughout this thesis. Then, we delve into the main community detection algorithm employed in our research: the hierarchical stochastic block model.

2.3.1 Bipartite networks to represent patients clinical data

The benefits of using networks to capture complex phenomena have been demonstrated across various healthcare domains. Networks have been used to represent relationships between cellular components across organs (Barabási, 2013) and to model complex multimorbidity patterns in patients (Kalgotra et al., 2017, 2020). While networks can effectively encode complex relationships, it is fundamental to determine the specific aspects of the data to highlight. Our first step was to define the type of network that efficiently represents patients and their characteristics to meet our research goals.

Considering the nature of our dataset, where clinical features occur in the context of individual patients, we aim to explicitly represent the relationship between patients and their specific characteristics. An affinity or similarity graph, while common, would have been less ideal as it emphasizes pairwise relationships rather than encoding patients' characteristics individually. Thus, we chose to maintain distinct nodes for patient and clinical features (e.g. morbidities and clinical admission data in Chapter 4 and severity of illness scores in Chapter 5). The differentiation between the two types of nodes led us to propose a bipartite network to encode the relationships between patients and clinical features. In this type of networks, nodes of one type are connected only to nodes of the other type. Nodes of the same type are not connected to each other. Figures 4.6 and 5.3 show the bipartite networks we use in our research.

While not exclusive to the use of a bipartite network, we identify several benefits throughout our work. The split between patients and features preserves important characteristics of our data, such as the number of co-occuring morbidities in a patient, represented by the patient node degree (Chapter 4). Additionally, the community inference process is simplified as the number of links between nodes is limited by not having connections between nodes of the same type. The final networks of patients and their features as nodes are presented in Sections 4.2.4 and 5.2.4.

2.3.2 Stochastic block modelling

Stochastic block models (SBMs) are generative models used for community detection in networks (Holland et al., 1983). In these models, the nodes in the network are divided into groups (i.e. blocks) such that the probability of a node connecting to other nodes depends only on their block membership. In this way, SBMs assumed a conditional independence between nodes in the same block, similar to other approaches such as LCA discussed in 2.2.2.

A common formalization of the SBM model can be found in literature (Karrer and Newman, 2011), and starts by modelling a network G with N nodes in terms of the Adjacency matrix A, of size N x N, with $A_{i,j}$ 1 if a link between nodes i and j exists and zero otherwise. The model goes to divide the nodes into B blocks to group nodes with similar patterns of connectivity to the rest of the network. Considering these B blocks a new adjacency matrix E is defined to represent the count of links or edges between elements in each pair of blocks. Matrix E is symmetric, with $e_{r,s}$ as the expected value of the $A_{i,j}$ for nodes i and j assigned respectively to blocks r and

$s \in B$. The probability of G can then be defined in terms of e and B as $P(G|e,B)$. The generative model is fitted by maximizing this probability in two steps: first with respect to the parameters $e_{r,s}$ and then with respect to the number of blocks B (Karrer and Newman, 2011). It is important to note that up to this point G can be directed or undirected, the number of blocks B is a model hyperparameter and e represents the expected number of links between nodes.

The standard SBM described so far has several drawbacks extensively discussed in the literature (Karrer and Newman, 2011; Peixoto, 2014b, 2017; Funke and Becker, 2019). Firstly, there is no principled mechanism to define the number B of blocks. Secondly, the model cannot identify statistically significant blocks with size smaller than $O(\sqrt{N})$, known as the resolution limit. Thirdly, it fails to capture networks with heterogeneous node degrees such as social networks and hierarchical structures. Finally, the likelihood maximization process is computationally inefficient.

The degree-corrected stochastic block model (DC-SBM) (Karrer and Newman, 2011) is a SBM variant proposed to better capture real-world networks. The core difference from the traditional SBM is that the DC-SBM includes the node degree as an additional parameter, hence the name degree-corrected. The explicit inclusion of the degree of nodes allows the DC-SBM to capture network structures showing a heterogeneous node degree. In this model, the node degree is estimated following a likelihood maximization as in the standard SBM for B and $e_{r,s}$. Given this new parameter, the probability of G is now defined as $P(G|k,e,B)$. While this model shows improved performance, it still lacks a principled mechanism to define the network blocks and suffers from the resolution limitation (Peixoto, 2017; Funke and Becker, 2019).

In the next section, we introduce the hierarchical stochastic block model (hSBM) (Peixoto, 2017), a non-parametric extension of the DC-SBM proposed to further address the limitations of SBMs. We use this method as an alternative to traditional clustering in Chapters 4 and 5.

2.3.3 Hierachical stochastic block modelling

The hierarchical stochastic block model (hSBM) (Peixoto, 2017) is a variant of the DC-SBM to further address two main limitations of SBMs: avoid the resolution limitation and to achieve a principled approach to infer the model parameters from the statistical evidence available in the data. Their approach proposes a Bayesian alternative to estimate the probability of observing the network G given the parameters of

the DC-SBM in order to sampling the model from a posterior distribution rather than finding the most likely partition as in the DC-SBM.

The first modification to the DC-SBM model is that in the hSBM the node degree k is fixed exactly to the observed network rather than only in expectation. This hard constrain simplifies the inference process as any network configuration that does not meet the exact node degree does not need to be considered. This constraint leads to the microcanonical degree-corrected SBM (Peixoto, 2017) which is the foundation for the final hSBM model.

The second step is to propose a Bayesian model for the $P(G|k,e,B)$ to infer a non-parametric framework to infer the parameters from the available evidence from the data. Equation 2.6 shows this formulation, with the network G represented by its adjacency matrix A and $P(k|e,b)$, $P(e|b)$, and $P(b)$ prior probabilities (Peixoto, 2017). In the following we present a brief description of the hSBM inference process, a complete description can be found in the hSBM presentation paper (Peixoto, 2017).

$$P(A,k,e,b) = P(A|k,e,b)P(k|e,b)P(e|b)P(b) \tag{2.6}$$

The process begins with the estimation of parameters in equation 2.6 based on evidence from the observed network A. Using the microcanonical formulation and assuming no-empty blocks b, priors for blocks $P(b)$ and the degree sequence $P(k|e,b)$, are proposed based on the characteristics of the observed network A. A key point in the derivation of the hSBM comes when defining the prior for the number of edge counts e. To allow for the discovery of small communities, a prior is proposed by interpreting the matrix E as a new adjacency matrix A' for a network with B nodes and $e_{r,s}$ edges. This hyperprior connects the SBM for $P(A)$ to a second SBM for $P(A')$. The procedure is repeated until a level l where the number of blocks $B_l = 1$.

One of the characteristics of hSBM is its robustness to overfitting (Peixoto, 2017). An intuitive understanding of this characteristic can be achieved by examining an alternative representation of equation 2.6 taken from the information theory domain (Rissanen, 1978).

$$P(A,k,e,b) = 2^{-\Sigma} \tag{2.7}$$

Where

$$\Sigma = -\log_2 P(A,k,e,b) \tag{2.8}$$

$$= -\log_2 P(A|k,e,b) + -\log_2 P(k,e,b) \tag{2.9}$$

In equation 2.9, Σ represents the model's description length, indicating the bits necessary to describe it (Rissanen, 1978). Notably, by maximizing equation 2.6, one is automatically minimizing the description length in equation 2.9. It is useful to note that the description length can be interpreted as the sum of two terms: one for the description of the model given the network parameters (first term on the right-hand side of equation 2.9) and another for the description of the model parameters (second term on the right-hand side of equation 2.9). Therefore, when achieving the model with the minimum description length (MDL) we are finding the simplest model that maximizes the probability of observing A. The MDL is useful for comparing models with different structural components like different numbers of blocks, and is the criteria we use in our model selection.

Model selection

While the selection between models can be done following the MDL criterion, we need to first obtain optimal partitions of the network. To achieve this, we utilize an efficient Monte Carlo heuristic (EMC) proposed for efficient inference of SBM models (Peixoto, 2014a). Starting from a given partition, the algorithm proposes a mechanism to find the most stable block structure, as the one with the lowest MDL from the possible partitions satisfying the microcanonical SBM posterior.

The EMC consists of a process in which the block membership of each node is randomly modified and said move is accepted with a probability given as a function of the entropy loss induced by the move. The probability is such that the process is ergodic (i.e. all possible partitions are accessible) and all moves are reversible given sufficiently long running time (Peixoto, 2014a). The EMC proposes an optimized membership move strategy which is further improved by allowing it to propose moves of groups of nodes, or agglomerative moves, to improve its speed. The ergodic characteristic of the EMC and its optimized speed thanks to the agglomerative moves makes it our choice for a first exploratory exploration of the solution space and identify good candidates for the optimal solution.

Although the EMC can effectively explore all the space of possible network partitions, it might still be possible to further refine the model previously obtained. For this purpose, we employ a second heuristic, the merge-split Markov chain Monte Carlo (merge-split MCMC) (Peixoto, 2020). The merge-split MCMC is effectively an extension of the EMC in that it has an agglomerative move strategy (the merge) but adds a split strategy. By doing this, the merge-split MCMC can better refine partitions further

avoiding local minima (Peixoto, 2020).

In Chapters 4 and 5, we apply these two algorithms sequentially to ensure the automatic inference of stable and optimal models. we run this approach 100 times, inferring 100 different solutions and choosing the one with the best MDL.

Priors for the inference process in the hSBM

In the original hSBM formulation (Peixoto, 2017), the priors for $P(k|e,b)$, $P(e|b)$, and $P(b)$ are chosen in an non-informative way. The rationale behind this choice is to prevent biases in the observed graph G from influencing the block structure at higher levels. However, naively using uninformative priors could lead to sub-optimal results, such as allowing undesirable partitions, for example, partitions with empty blocks (Peixoto, 2014b). To mitigate this, a series of hyperpriors are proposed to incorporate relevant information to every hierarchical level. In the following we will shortly exemplify how this is achieved, an extensive explanation can be found in relevant literature (Peixoto, 2014a, 2017).

An example of this approach is the prior for the probability of a block assignment b given a block partition B. A completely uninformative prior would be $P(b|B) = B^{-N}$. Although uninformative, this prior assumes that the blocks b_i would roughly have similar sizes and it allows for empty blocks. The proposed solution is to incorporate a parametric distribution for the probability of $P(b)$ conditioned to the group sizes n_r and then to define a prior for the probability for the number of blocks B. In this way the prior for $P(b|B)$ is defined as (Peixoto, 2017).

$$P(b|B) = P(b|n)P(n|B)P(B) = \frac{\prod_r n_r!}{N!}\binom{N-1}{B-1}^{-1}\frac{1}{N} \tag{2.10}$$

The terms in the prior in equation 2.10 can be understood as follows. A first term for the probability of a block assignment b conditional on the size of groups $P(b|n)$. A second term for the group size conditional to the block partition $P(n|B)$. This is expressed as the number of histograms with B bins with elements summing up to N. Finally a third prior for the number of groups B, in this case simply $1/N$. The last two terms are defined assuming only nonempty blocks. A detailed description of the definition of these prior can be found in the literature (Peixoto, 2017, 2019).

Another example is the in the priors for the node degree. Interestingly in this case the uninformative hyperpriors are hierarchical, as they are defined in terms of the total number of edge counts which remains constant for all of the hierarchical models of

the hSBM. Since E is constant the SBMs become denser in terms of edge count. This density pushes the model to eventually cluster all blocks into one, marking the end of the inference process.

While non-informative priors are useful in cases when we want to explore the data, there are cases where it is not desirable. For example in cases where we have relevant domain information as in the healthcare domain. Mechanisms to address this can be found in literature, for example, to inform the model about the expected probability on an edge (Peixoto, 2021). Although we do not use these tools in our current work, exploring this avenue could be beneficial.

Understanding the hSBM in our work

When applied to our work, it is important to note some aspects of the results yielded by the hSBM. In particular, we would like to comment on the relationship between patients and features in the context of our bipartite network and then on the choice of priors.

While we force a bipartite split between patients and clinical features, this does not mean that the clustering inference is done independently. On the contrary, the patient and feature clusters are inferred simultaneously. This means that changes in the observed conditions of patients and/or features will have implications for the block structure inferred. Strictly speaking, both patients and clinical features only interact at the highest hierarchical level. In other words, in the adjacency matrix of every level, but the highest one, we will observe values greater than zero only between blocks of the same type (i.e. patients and features). But thanks to hard constraints on the node degree of the microcanonical hSBM and the hierarchical hyperpriors for the edge counts these two sets of seemingly disjoint groups of clusters are linked.

In terms of the uninformative priors used, it is important to note that non-informative does not necessarily mean uniform, as presented in section 2.3.3. This flexibility allows hSBM to unveil complex block structures with diverse parameters. For instance, in Chapter 4 the hSBM unveils both small (under 5% of the population) and large (over 20% of the population) clusters and of varying edge count (i.e. clusters of nodes with exactly 1 node and others with more than 5 nodes).

In conclusion, the application of hSBM in our research provides a framework for understanding the relationships within our bipartite network. By simultaneously clustering patients and clinical features, we achieve a comprehensive model that reflects the complexity of our data.

2.4 Prediction models for mortality

Predicting ICU patient mortality is crucial for patient assessment, resource allocation, and medical care evaluation (Lin et al., 2019). Leveraging available clinical data, machine learning has been widely explored for this purpose. In this section, we focus on tree-based models, specifically Random Forest (Breiman, 2001) and XGBoost (Chen and Guestrin, 2016) due to their widespread use in ICU related predictions (Keuning et al., 2020; Ishwaran et al., 2008; Kong et al., 2020; El-Manzalawy et al., 2021). Specifically, in Chapter 5, employ these methods in the development of our proposed mortality predictive model, hXG-SAPS.

These models are based on the concept of a **decision tree**. In a dataset with N observations, each having features X and possible outcomes Y, a decision tree $f(x)$ is a tree-like structure used to predict outcomes y. Each tree node represents a decision based on a specific feature x_i in X and a feature-specific threshold, branches represent the outcome of that decision, and leaf nodes provide the final decision or prediction (Murphy, 2012). The model is trained to learn the sequence in which to examine features and their specific thresholds to maximize the probability of observing the outcomes Y given an observation represented by a set of features X.

A major limitation of decision trees is their instability, leading to varying predictions based on the training data (Murphy, 2012). **Random Forest** (RF) addresses this by leveraging multiple decision trees to create a more stable model. As presented in equation 2.11, predictions are achieved by averaging the outcomes of M trees. Each tree is trained on a unique training set, built on different subsets of features and randomly sampling observations from the main training set (Breiman, 2001).

$$f(x) = \sum_{m \in M} \frac{1}{M} \cdot f_m(x) \tag{2.11}$$

XGBoost (Chen and Guestrin, 2016) is an alternative approach to address the limitation of decision trees, considering multiple trees to achieve an outcome. Unlike RF, where trees are trained independently, XGBoost sequentially adds trees to correct the model's current performance. This is achieved by minimizing the objective function L, over T steps:

$$L^t = \sum_{n \in N} l(y_i, \hat{y}_i^{t-1} + f_t(x_i)) + \Omega(f_t) \tag{2.12}$$

Here, $\hat{y}_i^{t-1}$ represents the prediction for y_i in step $(t-1)$, l is a function measuring

the difference between y_i and its prediction, f_t denotes the new tree to add to the model in step t and Ω is a function penalizing the complexity of f_t. XGBoost is known for its scalability, robustness to overfitting and high performance (Chen and Guestrin, 2016).

2.5 Evaluation metrics for model performance

The performance assessment of various clustering and prediction models is key to our work. In this section, we introduce the metrics used throughout this thesis for this purpose. We begin with clustering metrics used in our research on multimorbidity profiles. Then, we present metrics used to evaluate the quality of predictive models, relevant to our work on mortality prediction. Additional discussions of these metrics in the context of medical usage can be found later in Chapters 4 and 5.

2.5.1 Metrics to evaluate clustering models

The evaluation of clustering models is challenging, mainly due to the lack of a ground truth (Murphy, 2012). Various metrics exist to help assess the quality of a model and enable comparisons between models. In this section, we introduce the metrics we use to support our work on multimorbidity profile identification in Chapter 4.

Model selection

In addition to distortion, BIC and AIC scores discussed in Section 2.2.4, we employ the Calinski-Harabasz (CH) score for model selections in Chapter 4. The CH score, originally introduced in the context of hierarchical clustering (Caliński and Harabasz, 1974a), is presented in equation 2.13 where $BGSS$ (between-group sum of squares) is a measurement of the distance between clusters, $WGSS$ (within-group sum of squares) a measurement of the dispersion of clusters, N the number of observations and K the number of clusters.

$$CH = \frac{BGSS}{WGSS} \cdot \frac{(N-K)}{(K-1)} \tag{2.13}$$

The CH score measures the extent to which observations are grouped into distinct clusters while minimizing the dispersion within each cluster. A high CH score indicates a more defined clustering structure, making it a valuable tool for comparing and selecting different models (Caliński and Harabasz, 1974a). We use this metric in Section 4.2.3 to measure the clustering error for the elbow calculation in agglomerative

clustering. Additionally, we use it as a tool to compare the quality across all of our models in Section 4.3.1.

Model stability

Another approach to evaluating a clustering model is to compare it to another model. The Mutual Information index (MI) is used for this purpose, quantifying the agreement between two partitions. MI measures the probability that observations are assigned to the same clusters in two different partitions U and V. Considering $P(i,j)$ as the probability that a randomly chosen object falls simultaneously in clusters $i \in U$ and $j \in V$, and $P(i)$ and $P(j)$ the probabilities that an object is assigned to cluster i and j respectively, then the MI is calculated as follows:

$$MI(U,V) = \sum_{i \in U} \sum_{j \in V} P(i,j) \log \left(\frac{P(i,j)}{P(i)P(j)} \right) \tag{2.14}$$

We use a normalized MI, which ranges between 0 and 1, with a higher MI representing a higher level of agreement between the partitions. In Section 4.2.2.2, we use the MI to compare our selected K-modes and LCA models against the remaining top nine performing models in an effort to further validate our models. By doing this, we aim to ensure that our models represent stable clustering structures and are not a result of overfitting or local minima.

Similarity between observations

A common task in clustering is the quantification of similarity between observations in a dataset. The Jaccard coefficient or similarity index, is a measure commonly used to this end (Murphy, 2012). Figure 2.15 presents the this similarity index considering two observations X_i and X_j, each composed by a set of categorical features.

$$J(X_i, X_j) = \frac{|X_i \cap X_j|}{|X_i \cup X_j|} \tag{2.15}$$

The similarity between these two observations is defined as the proportion of shared elements between the observations. This similarity index is defined between 0 and 1, with 0 meaning no similarity and 1 complete similarity.

2.5.2 Metrics to evaluate predictive models

Evaluating predictive models, like Random Forest or XGBoost, is crucial for our work on patient severity of illness in Section 5. Specifically, we assess our models on three key categories: their ability to predict patient outcomes, their capacity to rank patients by the probability of presenting an outcome, and their effective estimation of outcome prevalence at a population level.

To guide our presentation, let us consider a dataset of observations, each represented by a features vector (e.g. patients' age or gender) and a binary label denoting an associated outcome (e.g. survived or not). We define P as the count of positive labels and N as the count of negative labels (e.g. values 1 and 0 respectively). Next, we present a number of performance metrics used in the healthcare domain.

Model's ability to predict patient outcome

An intuitive way to evaluate a predictive model is by counting the number of correct predictions. Specifically, we would want to determine the number of observation correctly classified in the positive and negative classes, true positives (TP) and true negatives (TN) respectively. Conversely, we can count the number of observations incorrectly classified in the positive and negative classes, or false positive (FP) and false negatives (FN). Using these concepts, we can define the model's **Accuracy** (equation 2.16) as the probability of correctly classifying an observations, **Sensitivity** (equation 2.17) as the probability of correctly identifying a positive observation, and **Specificity** (equation 2.18) as the probability of correctly classify a negative observations.

$$Accuracy = \frac{(TP+TN)}{P+N} \tag{2.16}$$

$$Sensitivity,\ True\ Positive\ Rate\ (TPR) = \frac{TP}{P} \tag{2.17}$$

$$Specificity,\ True\ Negative\ Rate\ (TNR) = \frac{TN}{N} \tag{2.18}$$

While widely used, these metrics present limitations for objectively evaluating models. Firstly, As our models are inherently tied to the training data, their assessment may not generalize to new datasets. To mitigate this, we conduct a *bootstrapped* evaluation (Efron, 1992) of our models' accuracy, sensitivity and specificity in Chapter 5. That is, we compute these metrics on randomly sampled testing datasets from our

primary test dataset. The result is a distribution for the metrics rather than a single value, providing a more comprehensive evaluation of our models.

A second limitation applies to regression models for binary prediction, such as Random Forest and XGboost. These models produce predictions based on a threshold used to binarize their outcomes. The choice of this threshold can significantly impact the accuracy, sensitivity, and specificity calculation. To address this, we employ Youden's J statistic (Youden, 1950) in Chapter 5. This metric selects the optimal threshold for binary classification by minimizing the trade-off between specificity and sensitivity.

Model's capacity to rank patients by the probability of presenting an outcome

Instead of finding an optimal binarization, we could look at the performance of a model using various thresholds to assess it independently from any specific one. This is commonly done by plotting the TPR versus the FPR for a model at different thresholds. This plot is known as the **receiver operating characteristic** or **ROC** curve (Fawcett, 2006).

The quality of a model can be summarized as the areas under the ROC curve, known as **AUC**. The AUC ranges between 0.5 and 1.0, with 0.5 indicating a model with no better performance than chance and 1.0 indicating a classifier with perfect predictive power. Notably, it has been proved that the AUC is equal to the probability that a randomly chosen positive observation has a higher prediction value than a randomly chosen negative one (Fawcett, 2006). This characteristic makes the AUC a widely used metric for evaluating the power of a model to discriminative between positive and negative observations.

Model's ability to predict outcome prevalence at a population level

An alternative method to assess a model without relying on arbitrary thresholds is to directly evaluate its output as an approximation to probability. Calibration curves (Hartmann et al., 2002) serve this purpose by comparing the prediction of a binary classifier to the observed frequency of the outcome. This curve plots the frequency of the positive label on the vertical axis, conditioned to the predicted probability on the horizontal axis. Figure 2.2 shows a calibration curve reported Table 5.3 of Chapter 5. To construct the curve, observations were grouped based on their predicted mortality into consecutive bins representing 10% ranges (e.g., 0-10%, 10%-20%). On the vertical axis, the frequency of observed mortality for patients in each bin is plotted.

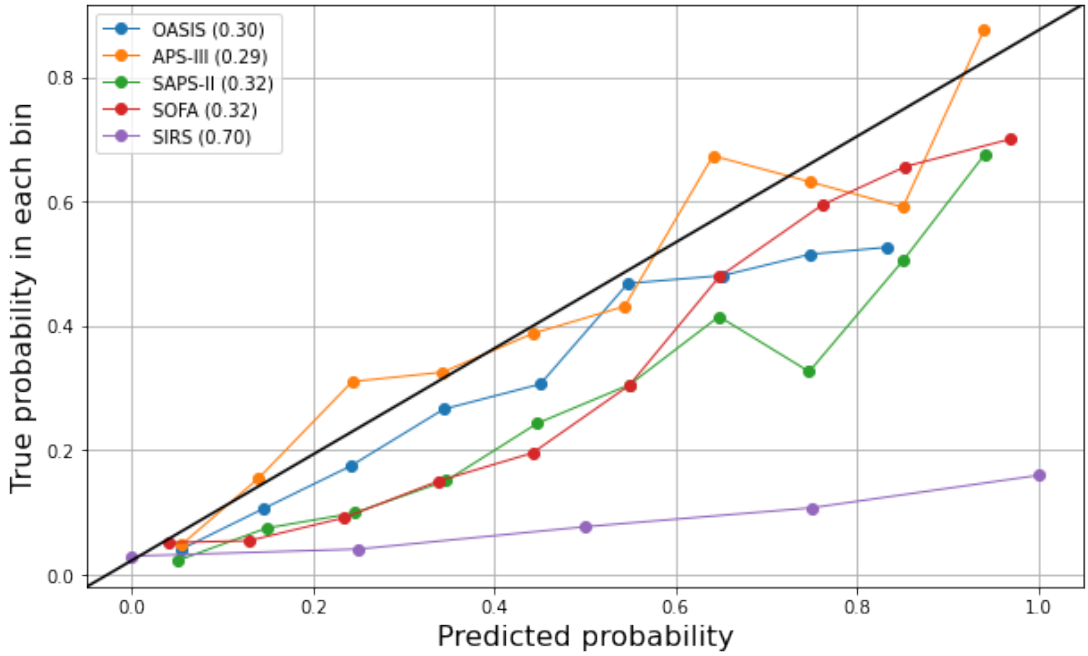

Figure 2.2: Calibration curves and RSME for severity score used as mortality predictor. Higher RSME reflects a higher correlation with the line of perfect calibration in black.

A perfectly calibrated model would show a 45-degree line, indicating that on average, the predicted probability matches the observed frequency of the outcome. Deviation from this perfect calibration line can be used as an indication of the output calibration of a model. Following relevant medical literature (El-Manzalawy et al., 2021; Walsh et al., 2017), we use root square mean error (RSME) between the calibration curve of a model and this perfect calibration line as a calibration metric (See Section 5.2.5). RSME values closer to 0 indicate better calibration performance.

A similar metric used in the healthcare domain is the Observe-to-Expected mortality ratio (O/E) (Strand et al., 2009; Zimmerman et al., 2006; Moreno et al., 2005). Calculated as the total number of observed deaths divided by the sum of the predicted probabilities of death, O/E provides a general view of the model at an aggregated level without the need for binarization. It is important to note that O/E emphasizes the expected deaths, being more lenient about the quality of the predictions than RSME and the ranking of patients than AUC.

2.6 Conclusions

In this section, we presented key concepts from the healthcare and medical domains, machine learning, and network science that are essential for the presentation of our research in the upcoming chapters. Specifically, we have introduced multimorbidity and its connection with electronic health records via the ICD-9 coding system, and severity of illness scores, which are tools used to guide medical decisions in the ICU. Additionally, we discussed technical elements spanning machine learning and network science that are applied in our work on multimorbidity profile identification and severity of illness work.

Importantly, we have highlighted some of the limitations of current clustering algorithms. These include the complexities in the model selection, challenges in handling missing and sparse data, and reliance on statistical assumptions about the data. These factors affect both the performance and adoption of machine learning in the ICU. At the same time, we have emphasized the advantages of the non-parametric approach for community detection in networks as an alternative to current clustering methods. In the next chapter, we introduce our datasets, highlighting the characteristics and challenges they pose for our work.

Chapter 3

The MIMIC-III healthcare dataset

Data plays a crucial role in our research. At a fundamental level, the available data will define the depth of our patient analysis, the techniques we employ, and the reliability of our findings. More broadly, methodological aspects of the data collection, such as its geographical origin or resources involved, can limit the validation and generalizability of results. In this section, we focus on our dataset, aiming to provide context for our research, presenting the steps followed to construct the study cohorts and highlighting potential challenges this may present for the application of machine learning.

We start by briefly introducing the third version of the Medical Information Mart for Intensive Care (MIMIC-III), the primary data repository used in our research. We then present the retrospective calculation of sepsis, multimorbidities, and severity of illness scores based on the MIMIC-III data. Finally, we discuss the process we used to build the patient cohorts used in Chapter 4 for multimorbidity profile identification and in Chapter 5 for mortality prediction.

3.1 Medical Information Mart for Intensive Care III

MIMIC-III is a large, anonymised, and publicly available dataset of over 50,000 hospital admissions to the critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012 (Johnson et al., 2018). We use it for three main reasons. First, the database is comprehensive, covering more than a decade of patient admissions with detailed information about individual patient demographics and care. Secondly, it has been widely validated in healthcare research (Johnson et al., 2016d, 2018), particularly in multimorbidity profile identification (Zador et al., 2019) and mortality prediction (El-Manzalawy et al., 2021) research. Finally, MIMIC-III is actively maintained by

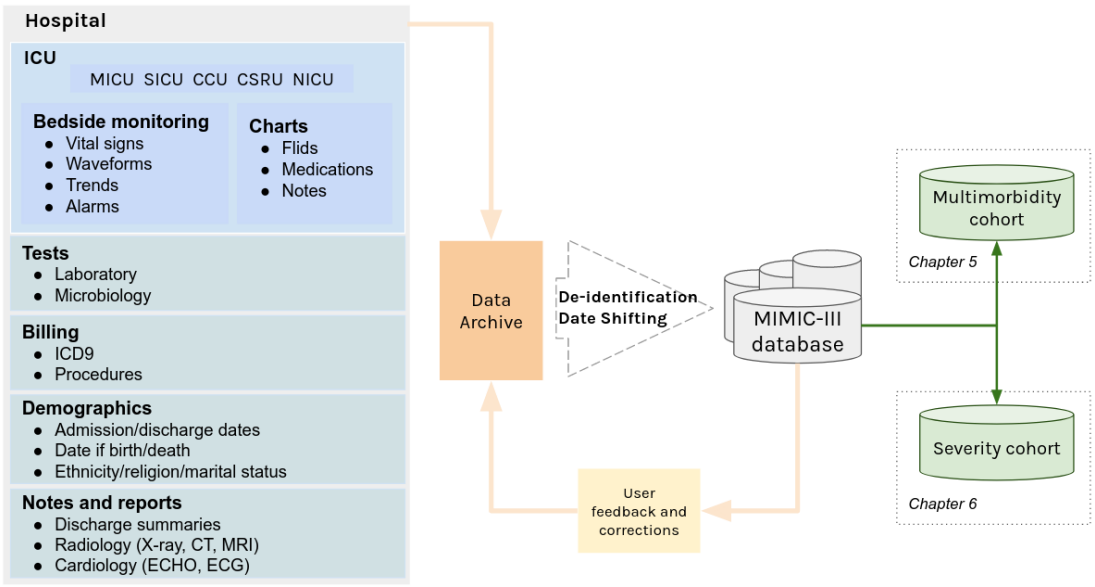

Figure 3.1: Overview of the construction of the MIMIC-III (Johnson et al., 2016d) and our study cohorts. It starts with the collection of hospital data and its processing until the MIMIC-III data is completed. From this database, we construct the cohorts for our multimorbidity profiles and patients' severity studies in Chapters 5 and Chapter 4

a specialized team, ensuring the database´s accuracy and enabling the replication of research findings. For example, the team updates and curates the data, and provides computed metrics (e.g. sepsis, comorbidities, etc.).

Figure 3.1 presents an overview of the data collection, processing and storage involved in the creation of MIMIC-III. The process starts by collecting a wide variety of records related to the patients, such as bedside monitoring data, test results, medical notes, demographics and diagnoses. These come from five hospital units, namely the Medical Intensive Care Unit (MICU), the Surgical Intensive Care Unit (SICU), the Cardiac Care Unit (CCU), the Cardiac Surgery Recovery Unit (CSRU) and the Neonatal Intensive Care Unit (NICU). After collection, the data follows a de-identification process in which fields such as the patient's name, telephone number or address are removed. As part of this process dates are randomly shifted, and as a result, all patient stays occur at some time between the years 2100 and 2200. Finally, the date of birth for patients over 89 is modified to show them with ages of over 299 years. The processed data is stored in 26 tables that constitute the overall MIMIC-III database.

MIMIC-III provides a vast healthcare data resource suitable for various research purposes. However, its versatility comes at a cost, as it requires the extraction of ad-hoc subsets to support more specific research. To this end, we resorted to existing

literature to create relevant cohorts for our studies, aiming to ensure a fair comparison of results. In the next sections, we describe relevant aspects of the data used in our work. We begin by describing the available patient information relevant to our work. Then, we detail the process followed to create the study cohorts for our multimorbidity profile and patient severity work. Additionally, we provide an overview of patient comorbidities and of severity scores, fundamental to the rest of our work.

3.2 Describing patient's condition

MIMIC-III provides a comprehensive record of a patient´s hospital journey. This can include multiple admissions and readmissions to various hospital wards including the ICU, as shown in Figure 3.2. Here, the *patients* table contains patient characteristics, consistent across admissions, while *admissions* and *icustay* tables capture the hospitalization path. The *diagnoses* table summarizes the diagnoses identified during each patient's hospital stay. Additionally, these tables store the patients' demographic and admission type, used throughout our research.

3.2.1 Patient admission information

The *patients* table holds 46,520 unique patients and their basic information like gender, date of birth and death. Each patient is associated with one or more admissions recorded in table *admissions*, containing 58,976 unique records. These records include demographic and administrative data such as ethnicity, diagnosis at admission, admission type, discharge information and insurance type. Diagnosis information in *admissions* is linked to the *diagnoses_icd* table, holding 651,047 diagnoses encoded using the ICD-9 classification code, discussed in Chapter 2.1.2. The patient's primary diagnosis at admission is recorded explicitly (as $seq\_num = 1$), and all others are considered secondary. Simultaneously, each admission is linked to one or more ICU admissions in table *icustays*. This table holds information about 61,532 ICU admissions, including the ICU unit of admission and discharge, and length of stays.

In Figure 3.3, we present a summary of the patient's information in MIMIC-III based on four relevant patient characteristics: age, gender, ethnicity and admission type. We consider these characteristics due to their relevance in ICU-related research, particularly in the identification of multimorbidity profiles (Zador et al., 2019) and risk assessment (El-Manzalawy et al., 2021). In yellow we present the prevalence of these

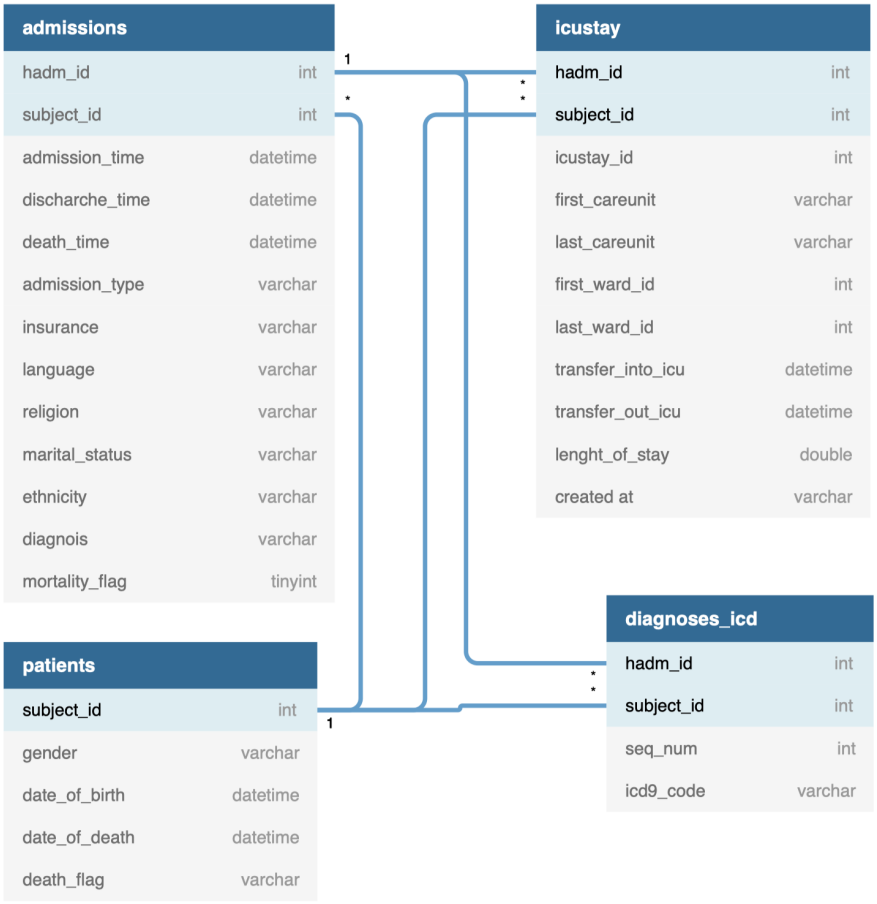

Figure 3.2: Tables holding patient information. While each patient has a unique identifier, *subject_id*, each patient can have multiple admissions to the hospital, *hadm_id*, and multiple ICU admissions during each hospital admission, *icustay_id*. It is worth noticing that multiple diagnoses can be associated with each hospital admission, in table *diagnoses_icd*, but no timestamp is provided. These tables also include the demographic information used in our multimorbidity profiles and mortality prediction work.

characteristics for all ICU admissions, in blue for the multimorbidity profile study cohort and in green for the mortality prediction cohort. Notably, the distribution of patients is consistent across our study cohorts.

Gender composition is slightly imbalanced across datasets, with approximately 59% male and 41% female patients. *White* ethnicity is the most common (over 70% for all cohorts), followed by *black* and a small number of *hispanic*, *asian* and *native american* patients. *Unknown* ethnicity refers to patients who did not report their ethnicity when asked and *other* corresponds to those reporting an ethnicity not categorized.

In terms of age, the database and study cohorts primarily consist of older patients, with an average age of close to 62 years old. Following existing approaches in the liter-

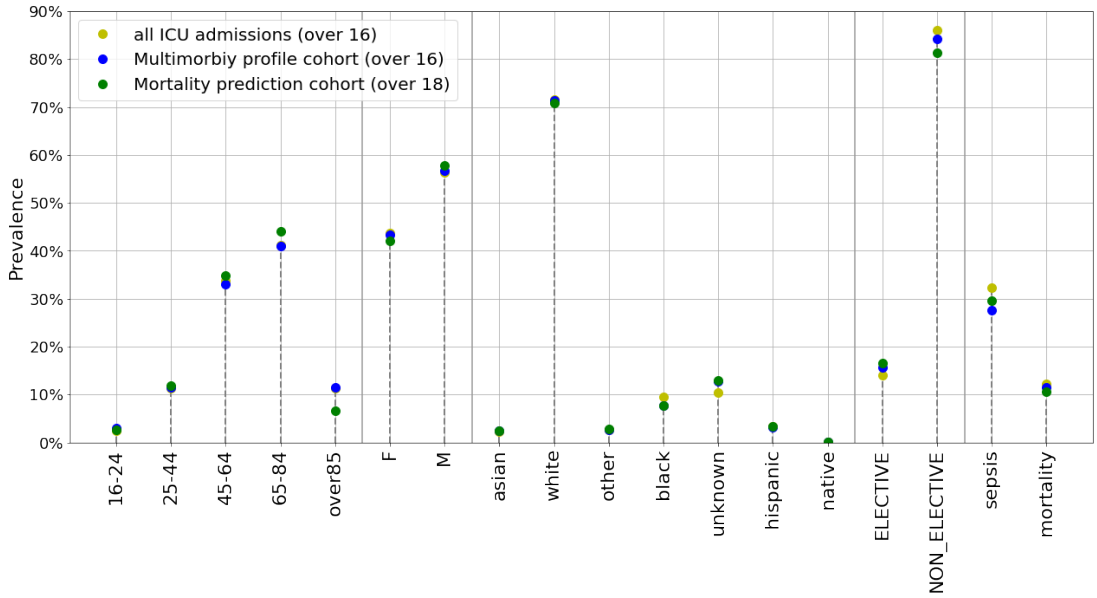

Figure 3.3: Patient gender, ethnicity, age, admission type, mortality, and sepsis are compared across all ICU admissions over 16 years old (yellow), multimorbidity profiles (blue), and mortality prediction (green) cohorts. Feature prevalence is similar in all samples, with an imbalance towards white patients over 45 years old with emergency admissions observed across cohorts. Minor age differences in our study cohorts result from relevant exclusion criteria for multimorbidity (patients over 16 years old) and mortality prediction (patients over 18 years old, included in the "16-24" age group for simplified visualization) studies.

ature (Geifman et al., 2013; Busija et al., 2019), we group patients into age categories to simplify the analysis and enhance the understanding of age-related relationships with the study variables. We consider patients in the following age groups [1]: 16-24, 25-44, 45-64, 65-84, and over 85. The cohort studies exhibit a similar age composition, with a majority of patients in the 65-84 and 45-64 age groups (around 40% and 35% of the sample respectively).

3.2.2 Patient mortality and sepsis

In-hospital mortality and sepsis are two widely studied patient outcomes in the literature related to multimorbidity profile identification and mortality risk assessment. Other than mortality, sepsis is of relevance as a critical factor in increasing the cost of

[1]Over 85 includes patients artificially recorded as being 299 years old and over to safeguard their anonymity.

care but, most importantly, increasing the probability of adverse patient outcomes such as death (Zador et al., 2019). In-hospital mortality is directly recorded in MIMIC-III in the table *admissions* under the field *"mortality_flag"*, as shown in Figure 3.2. From now on and for simplicity, we will refer to in-hospital mortality as mortality.

The prevalence of in-hospital mortality and sepsis is presented in Figure 3.3 for the whole ICU admissions of those over 16 years old (yellow), the multimorbidity profile described in the next section (blue), and the mortality prediction (green) study cohorts presented in Section 3.4. Notably, the prevalence of these outcomes is virtually the same in all cohorts, close to 28% for sepsis and 11.5% for mortality. It is worth noticing the imbalance of all of the cohorts in terms of outcomes.

As discussed in Section 2.1.3, sepsis is not always available in the MIMIC-III records, and it has to be computed based on available information. This is done following the definition discussed in Section 2.1.3, which specifies it as a combination of organ dysfunction simultaneously associated with a bacterial or fungal infection. Specifically, a patient is considered to have suffered from sepsis if: it is explicitly stated in the patient's history that there exists an infection (bacterial/fungal) *and* organ dysfunction, or if there exists an infection (bacterial/fungal) *and* the patient is under mechanical ventilation.

It is important to note that the Angus definition, while reasonable and widely used in research present, limitations such as their reliance on administrative data (Iwashyna et al., 2014). The Angus method is susceptible to miscalculation of sepsis due to missing data, inconsistency in coding practices or biases in coding based such as clinical judgment, financial incentives, or institutional priorities. For a detailed of the Angus definition of sepsis please refer to section 2.1.3.

3.3 Multimorbidiy profile identification study cohort

In this section we describe the process of creating a patient cohort for our multimorbidity profile study, presented in Chapter 4. To ensure the cohort made was relevant to our study and to conduct a fair comparison with existing research, we adopt the inclusion criteria described in a recent multimorbidity profile study (Zador et al., 2019), summarized in Figure 3.4. These criteria consider only the first admission of a patient, excluding patients under 16 years old or without recorded diagnoses.

Comorbidities and sepsis were derived from information in tables *patients*, *icustay*, *admissions*, and *diagnosis˙icd* (refer to Figure 3.2). Comorbidities are calculated

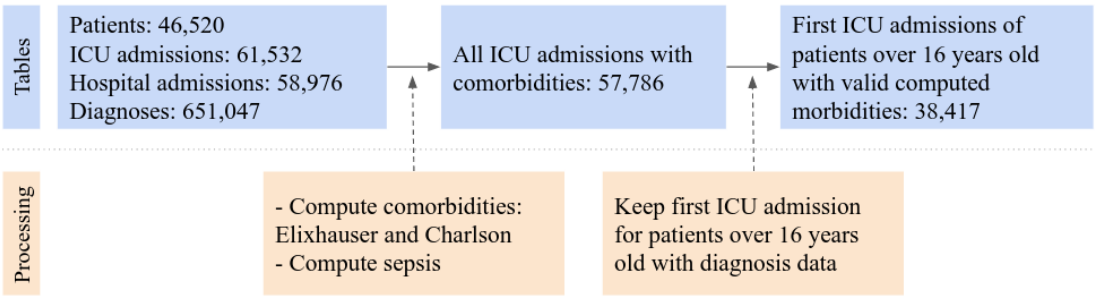

Figure 3.4: Process followed to obtain the multimorbidity cohort dataset. The final cohort contains 38,417 first admissions to ICU of patients over 16 years old.

considering only secondary diagnoses (seq_num ≥ 1 in table *diagnosis_icd*), as discussed in Section 2.1.2. Duplicate ICU admissions were removed, retaining only the first admission to avoid considering a patient multiple times in our cohort. It is important to note that while the unique hospital admissions are 58,976, only 57,786 of those have records in the *icustays* table. Finally, we excluded 180 patients for whom the comorbidity algorithms provide *undetermined* comorbidity values. Closer examination revealed that one patient lacked the primary ICD-9 code and the rest had no secondary ICD-9 code in the *diagnoses_icd* table. The final dataset includes 38,417 patient's first ICU admissions for patients of at least 16 years old. Following the above criterion we get a relevant cohort for our study, which interestingly follows closely the composition of the entire population (see Figure 3.3).

3.3.1 Comorbidity description

The link between patient comorbidities and medical outcomes has been widely documented (Barnett et al., 2012a; Prados-Torres et al., 2014). Although important, it is not always easy to identify patients' comorbidities, especially in the ICU context as the critical condition of patients makes it difficult to make extensive patient assessments. As presented in Section 2.1.2, in MIMIC-III comorbidities are calculated utilizing the Elixhauser comorbidity index (Elixhauser et al., 1998) and the Charlson comorbidity index (Charlson et al., 1987) from administrative data recorded separately. Both of these comorbidity indices are well established and offer an algorithmic way to compute patients' comorbidities based on the ICD-9 codes associated to their hospital stay. In particular, we use the SQL implementation of the Elixhauser and Charlson comorbidity index provided by the curators of the MIMIC database (Johnson et al., 2018).

Figure 3.5 summarizes the prevalence of Elixhauser comorbidities in our multimor-

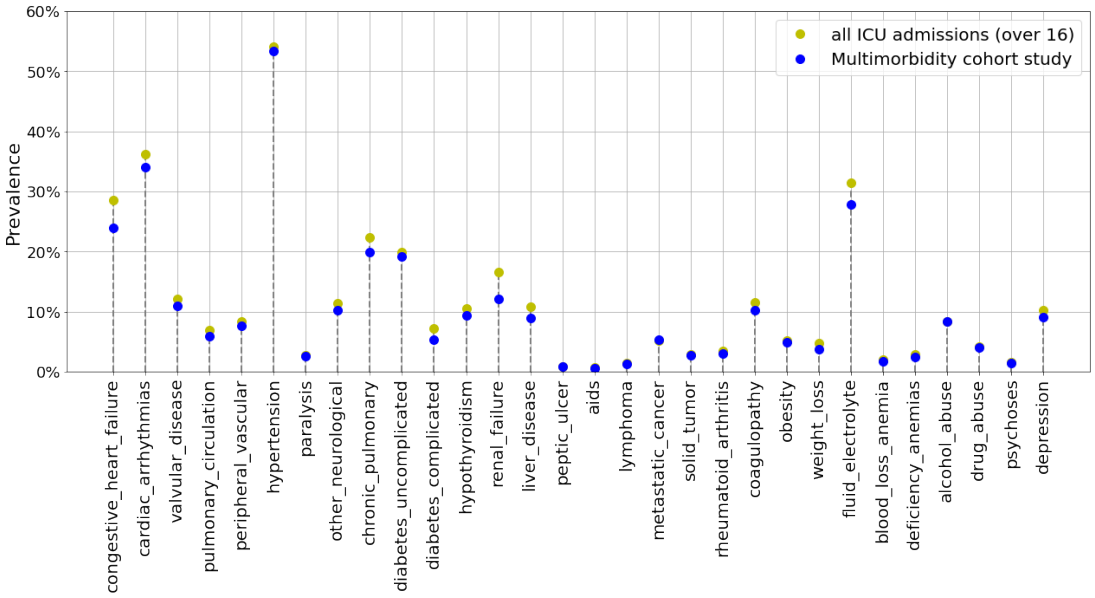

Figure 3.5: Prevalence of Elixhauser defined comorbidities in all ICU admissions and our study cohort. While comorbidity prevalence remains generally similar and low (under 10%), congestive heart failure, cardiac arrhythmia, hypertension and fluid electrolyte deficiency stand out with a prevalence over 20% in both datasets.

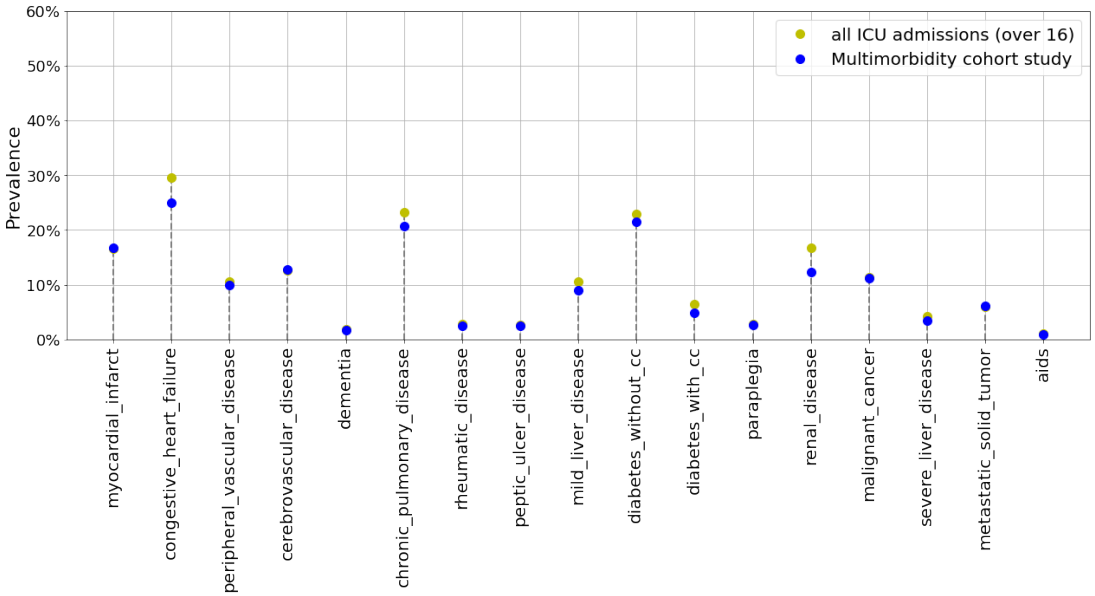

Figure 3.6: Prevalence of Elixhauser comorbidities in all ICU admissions and our study cohort as defined by the Charlson index. Comorbidity prevalence remains generally similar across datasets and under 30%. The main difference is a slightly higher prevalence of congestive heart failure, renal and chronic pulmonary disease in our cohort.

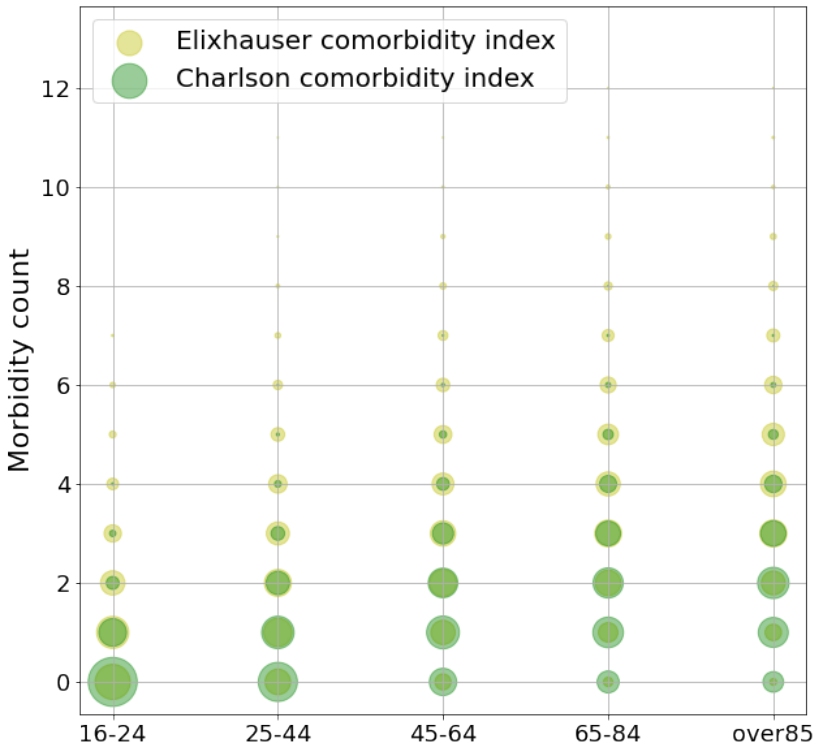

Figure 3.7: Multimorbidity counts by age group for Elixhauser (yellow) and Charlson (green) comorbidity indices in our multimorbidity profile cohort. Bubble sizes represent the frequency of patients with specific numbers of multimorbidities. For example, in the 16-24 age group, most patients have zero Elixhauser or Charlson comorbidities. Age shows a positive correlation with comorbidity count, consistent with existing literature.

bidity study cohort. It can be seen that the prevalence of morbidities remains virtually the same across datasets. In general, the prevalence of morbidities is low, with only hypertension showing a prevalence close to 50% in both samples. In fact, only congestive heart failure, cardiac arrhythmia, chronic pulmonary, fluid electrolytes deficiency and diabetes uncomplicated have a prevalence close to or higher than 20%. A second group of comorbidities can be identified with a prevalence close to 10%, namely valvular disease, chronic pulmonary, diabetes uncomplicated, and renal failure. All other comorbidities present a low prevalence, mostly under 5%.

Figure 3.6 presents the prevalence of Charlson comorbidities in all ICU admissions and our multimorbidity profile study cohort. Similarly to the Elixhauser case, a low prevalence of morbidities is observed across datasets. However, in this case, the prevalence is significantly lower, with no comorbidity showing a prevalence over 30%. The highest prevalence is of congestive heart failure (close to 27% in both datasets),

diabetes without complications and chronic pulmonary disease (both close to 20% in both datasets). Interestingly, congestive heart failure and diabetes are also defined in the Elixhauser index, and their prevalence remains similar in both datasets pointing to some level of consistency in the indices.

Multimorbidity count, a relevant aspect of multimorbidity linked to negative medical outcomes, is examined in Figure 3.7. This figure illustrates the number of comorbidities within age groups for both the Elixhauser and Charlson indices in our multimorbidity study cohort. Not surprisingly, multimorbidity count is rises with age, as supported by existing literature (Busija et al., 2019; Zador et al., 2019; Barnett et al., 2012b), with patients in the 16-24 age group showing the lowest count (close to zero), followed by the 25-44, 45-64, 65-84 and over 85 years group. It can be seen that regardless of the age group, patients would rarely present more than 4 or 5 comorbidities. Our results show this trend to be consistent across indices and datasets.

We have presented the study cohort used in Chapter 4 for our multimorbidity profile study. Our cohort resembles those extracted from MIMIC-III in relevant literature including demographic, admission type and Elixhauser comorbidities (Zador et al., 2019). Additionally, we have included the Charlson multimorbidity index with the goal of extending the validity of our work. Our analysis shows a dataset primarily consisting of white patients between 45 and 84 years old with non-elective admission. The cohort presents a significant imbalance in medical outcomes, with a low prevalence of mortality and sepsis. Finally, comorbidities exhibit low prevalence in the cohort across indices. This implies an imbalance between feature types, with dense demographic and admission-type features and sparse comorbidity features.

3.4 Mortality prediction study cohort

In this section, we describe the process of creating a patient cohort for our research on developing a severity score for mortality prediction. To ensure a fair comparison with existing state-of-the-art work, we follow the inclusion criteria described in OASIS+ (El-Manzalawy et al., 2021), as they tackle the same problem and use MIMIC-III. We include in our study cohort patients aged 18 to 90, with ICU stays lasting at least 24 hours and only first admissions when multiple ones are recorded. In terms of severity scores, we considered five due to their prevalence in the literature: SAPS-II, OASIS, APS-III, SOFA and SIRS. Figure 3.8 illustrates the process from the cohort extraction from the database to the final train and test datasets.

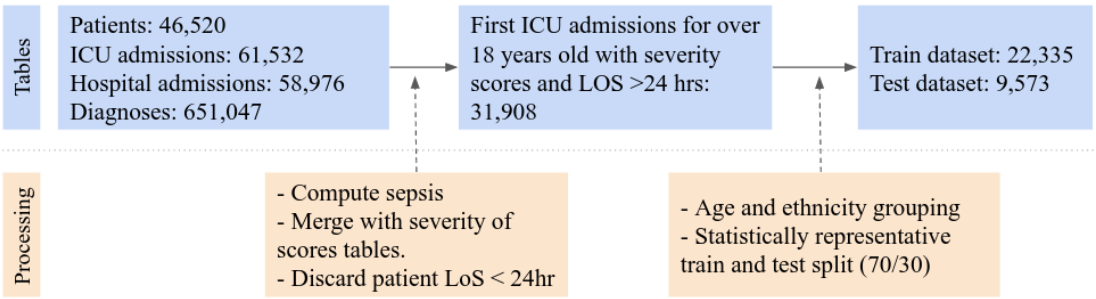

Figure 3.8: Construction of the mortality risk assessment cohort. The final cohort includes 31,908 patients, split into train and test datasets of 22,335 and 9,573 patients respectively.

We started by gathering patient information, hospital admission information and information on patients' clinical variables during their ICU stay from tables *patients*, *icustay*, *admissions* and *diagnosis_icd*. The query was restricted to patients between 18 and 90 years old with ICU stays of at least 24 hours to have enough data to compute the different scores. The final cohort included the first ICU admissions for patients following our inclusion criteria, adding up to 31,908 patients. The cohort is then processed to group patient age as presented in Section 3.2.1. Patients' ethnicity was categorized as White, Black, Asian, Hispanic, and Other, as defined in the MIMIC-III.

Finally, to support supervised learning, we randomly split the study cohort into train and test datasets (70% and 30% of the dataset respectively). To ensure the representativeness of these datasets we chose a partition with a statistically equal distribution of features (chi-square for categorical features, p-values ≤ 0.05). As shown in Table 3.1, both datasets exhibit virtually identical mean patient age (62.8 years old), length of stay (4.7 days), patient mortality (10.5% and 10.74%), and female/male composition, with the train set exhibiting 42.14% females and the test set 42.02%. The datasets exhibit similar average values for all severity scores studied.

3.4.1 Severity of illness scores

Severity scores are crucial for our work in mortality risk assessment. Of interest for our research are SAPS-II, OASIS, APS-III, SIRS and SARS, which we introduced in Section 2.1.4, describing each scoring system and the clinical variables each one considers. In this section, we present an overview of the scores within our study cohort. We describe the distribution of scores and briefly assess their quality for mortality prediction.

TRAIN				TEST			
	Total (n = 22,335)	Survivals (n =19,990)	Non-survivals (n =2,345)		Total (n = 9,573)	Survivals (n = 8,545)	Non-Survivals (n = 1,028)
Age (years)	62.76	62.10	68.33	Age (years)	62.79	62.10	68.50
LoS (days)	4.70	4.41	7.15	LoS (days)	4.70	4.40	7.19
Gender				Gender			
Female	9,413 (42.14%)	8,349 (41.77%)	1,064 (45.37%)	Female	4,023 (42.02%)	3,556 (41.61%)	467 (45.43%)
Male	12,922 (57.86%)	11,641 (58.23%)	1,281 (54.63%)	Male	5,550 (57.98%)	4,989 (58.39%)	561 (54.57%)
Age groups				Age groups			
16-24	606 (2.71%)	575 (2.88%)	31 (1.32%)	16-24	260 (2.72%)	252 (2.95%)	8 (0.78%)
25-44	2,655 (11.89%)	2,511 (12.56%)	144 (6.14%)	25-44	1,129 (11.79%)	1,062 (12.43%)	67 (6.52%)
45-64	7,794 (34.90%)	7,138 (35.71%)	656 (27.97%)	45-64	3,333 (34.82%)	3,060 (35.81%)	273 (26.56%)
65-84	9,797 (43.86%)	8,557 (42.81%)	1,240 (52.88%)	65-84	4,235 (44.24%)	3,668 (42.93%)	567 (55.16%)
Over 85	1,483 (6.64%)	1,209 (6.05%)	274 (11.68%)	Over 85	616 (6.43%)	503 (5.89%)	113 (10.99%)
Ethnicity				Ethnicity			
African	1,734 (7.76%)	1,590 (7.95%)	144 (6.14%)	African	742 (7.75%)	687 (8.04%)	55 (5.35%)
Asian	538 (2.41%)	475 (2.38%)	63 (2.69%)	Asian	225 (2.35%)	207 (2.42%)	18 (1.75%)
Hispanic	757 (3.39%)	707 (3.54%)	50 (2.13%)	Hispanic	304 (3.18%)	283 (3.31%)	21 (2.04%)
White	15,752 (70.53%)	14,192 (71.00%)	1,560 (66.52%)	White	6,843 (71.48%)	6,136 (71.81%)	707 (68.77%)
Others	3,554 (15.91%)	3,026 (15.14%)	528 (22.52%)	Others	1,459 (15.24%)	1,232 (14.42%)	227 (22.08%)
Severity of illness scores				Severity of illness scores			
APS-III	42.35	39.99	62.50	APS-III	42.13	39.70	62.27
OASIS	31.44	30.48	39.68	OASIS	31.39	30.48	38.98
SAPS-II	34.57	32.86	49.07	SAPS-II	34.47	32.75	48.78
SIRS	2.79	2.74	3.17	SIRS	2.80	2.76	3.15
SOFA	4.15	3.85	6.73	SOFA	4.11	3.81	6.56

Table 3.1: Summary information for train and test cohorts. Mean is reported for age, length of Stay (LoS), and severity scores. Percentage of the sample are indicated for sex, age group and ethnicity.

As discussed in Section 2.1.4, severity scores aim to quantify the condition of a patient using the information available from the first 24-48 hours. They map these clinical variables to subscores indicating their impact on the severity of a patient. For example, a patient presenting a high of its body temperature of 38ºC will receive a temperature subscore of **3 points** in SAPS-II (see Table 5.1), **2 points** in OASIS (see Table 2.1), and only **1 point** in SIRS (see Table A.2 in Appendix A.1). Subscores are them summed up to provide a final score, indicating the severity of the patient.

It is worth noticing that missing clinical variables require imputation before the calculation of severity scores. A common approach is to treat missing values as within normal clinical ranges (El-Manzalawy et al., 2021). For instance, if the patient's temperature in our previous example is unavailable, it will be assumed as normal. As a result, instead of the previously presented subscore, all scoring systems will assign a temperature subscore of **0 points**, reflecting a normal temperature. MIMIC-III implements this approach (Johnson et al., 2016c), and so it is the mechanism in our cohort.

This imputation mechanism is particularly relevant in our dataset due to the level of missing data and its impact in the computation of severity scores. As shown in figure 3.11, the level of imputation is significant for SAPS-II (pulmonary artery pressure imputation for 58,18% of the patients), SOFA (respiratory susbscore imputation for

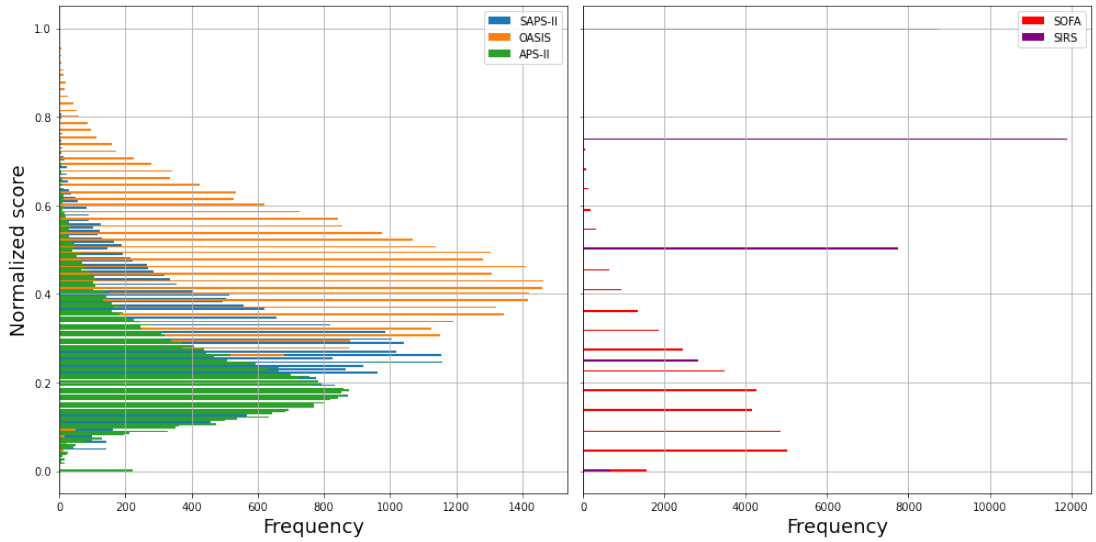

Figure 3.9: Distribution of the normalized severity scores for the study cohort. On the left, the distribution of SAPS-II, OASIS, and APS-III scores, and on the right for SOFA and SIRS.

48,96% of the patients), and APS-II (albumin levels in blood imputation for 63,76% of patients). Imputation is also present to a lesser extent for OASIS and SIRS. For OASIS, urine output is imputed for 4.32% of the patients, and body temperature is imputed for 3% of the patients.

To provide an overview of the scoring systems in our study cohort, Figure 3.9 displays the distribution of normalized severity of illness scores in our study. A first benefit of this view is that it facilitates comparisons among scoring systems with different ranges, as it is in our case. Additionally, the normalized values can served as proxies for mortality probability (El-Manzalawy et al., 2021) (with **0** as the lowest death probability and **1** the highest), providing a general trend of their predictions.

It is noticeable from the distribution of normalized scores, that SIRS generally predicts a higher probability of mortality, with a mean of 0.70. In contrast, all other scoring systems distributions show means below 0.5 with OASIS at 0.43, SAPS-II at 0.29, APS-III at 0.22, and SOFA at 0.18.

Figure 3.10 presents the AUROC performance of the severity scores as mortality predictors using their normalised score. Aligning with existing literature (El-Manzalawy et al., 2021; Strand et al., 2009), we see that SAPS-II (0.80) shows the best performance, followed by APS-III (0.78) and OASIS (0.76). As we discuss further in Chapter 5, these results reflect an excellent performance for these scores. In contrast, SOFA and SIRS show the poorest performance with 0.71 and 0.62 respectively.

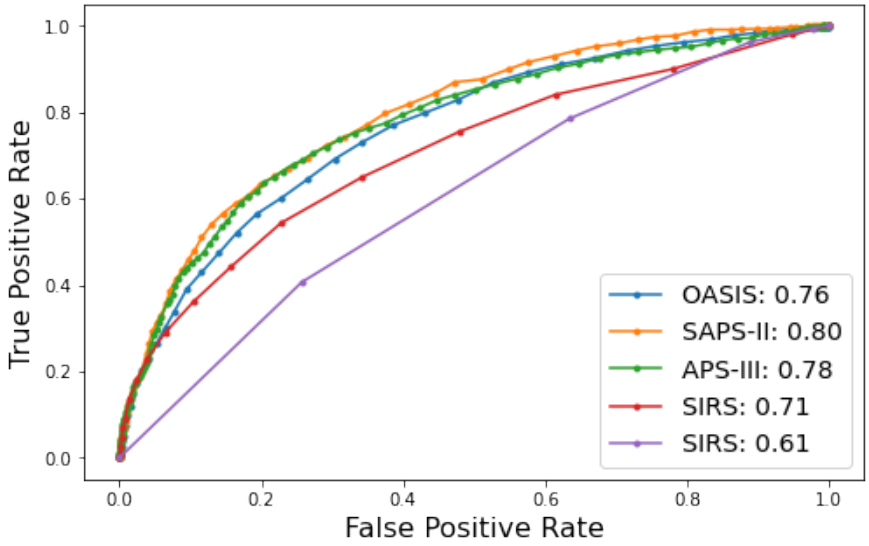

Figure 3.10: Receiver-Operator-Curve and Area under the curve for the prediction of mortality using the normalized scores for the test data of the study cohort
. SAPS-II shows higher performance indicating a high ability to rank patients in terms of their mortality risk, while SIRS shows the lowest discriminant power.

3.5 Conclusions

We have introduced MIMIC-III, a vast and publicly available data source and extensively used in healthcare research, especially in multimorbidity profile identification and mortality prediction. From MIMIC-III, we have constructed two study cohorts to support our upcoming work. These cohorts closely resemble those described in existing literature in terms of patient characteristics, comorbidities, and medical outcomes. This alignment contribute to the validity of our findings in the subsequent chapters.

However, we have encountered various of the data-related issues, as discussed in Chapter 1, in our cohorts. Firstly, a significant imbalance in medical outcomes (sepsis and mortality) and in the patients' demographic description. Secondly, the sparsity of morbidities is evident in comparison to other features such as demographics or severity of scores. Thirdly, we observe a significant level of missing data, as evident by the high level of imputed data in the computation of severity of illness scores. In the upcoming chapters, we use these study cohorts to explore our approach, aiming to enhance multimorbidity profiling and mortality risk assessment tasks while addressing these challenges.

Severity Score	Clinical variable	Missing data (%)
SAPS-II	Age (years old)	0,0
	Heart rate (beats/min)	1,01
	Systolic BP (mm Hg)	1,08
	Body Temperature (°C)	3,0
	Pulmonary artery pressure (P_aO_2, mm Hg/FiO2) *	58,12
	Urine output (L/d)	4,32
	Serum Urea Level (L/d)	0,6
	White Blood Cell count (WBC) (10^3/cu mm)	1,0
	Serum potassium (mmol/d)	0,45
	Serum sodium level (mmol/L)	0,55
	Serum bicarbonate level (mWq/L)	1,03
	Billirubin level (micro mol/L)	55,85
	Glasgow Coma Score (GCS)	1,05
	Chronic diseases	0
	Type of admission	0
SOFA	PaO2/FiO2 (mm Hg) [Respiratory]	48,96
	Platelets (10^3/mm^3) [Coagulation]	77,0
	BIllirubin (mg/dl) [Liver]	55,85
	Hypertension [Cardiovascular]	1,01
	Glasgow Coma Score (GCS) [CNS]	1,05
	Creatinine (mg/dL) [Renal]	0,14
SIRS	Body Temperature (°C)	3,0
	Heart rate (beats/min)	1,01
	Respiratory rate (breaths/min)	0,46
	White Blood Cell count (WBC) (10^3/dL)	1,0
* Only if patient is ventilated		

Severity Score	Clinical variable	Missing data (%)
APS-III	Heart rate (beats/min)	1,01
	Mean BP (mmHg)	1,02
	Body Temperature (°C)	3,0
	Respiratory rate (breaths/min)	1,10
	PAO2AADO2	43,86
	Hematocrit (%)	5,01
	White Blood Cell count (WBC) (10^9/L)	1,0
	Creatinine (mg/dL)	0,54
	Urine output (ml/24 hours)	4,32
	BUN (mg/dl)	0,60
	Sodium (mM/L)	0,55
	Albuminum (g/dl)	63,76
	BIllirubin (mg/dl)	55,85
	Glucose (mg/dl)	0,32
	Acidbase	32,75
	Glasgow Coma Score (GCS)	0,25
OASIS	Age (years)	0
	Pre-ICU (hours)	0
	Glasgow Coma Score (GCS)	1,05
	Heart rate (beats/min)	1,01
	Mean BP (mmHg)	1,02
	Respiratory rate (breaths/min)	1,10
	Body Temperature (°C)	3,0
	Urine output (cc/24 hours)	4,32
	Patient is ventilated	0
	Elective surgery	0

Figure 3.11: Proportion of patients with imputed subscores for each severity of illness given missing data in our cohort.

Chapter 4

Improving identification of multimorbidity profiles for ICU patients

The prevalence of multimorbidity in ICU patients is increasing as a consequence of a worldwide ageing population (Barnett et al., 2012a). Multimorbidity is a leading factor in adverse outcomes, such as sepsis and mortality, and it increases the cost of hospitalization (Reddy et al., 2020). Multimorbidity is particularly relevant in the ICU, where patients show the highest rates of acute and chronic conditions (Forte et al., 2019).

The relevance of multimorbidity has driven research in the identification of groups of patients sharing similar conditions profiles. Leveraging the abundance of ICU data in electronic health records, recent efforts have focused on machine learning to identify these profiles (Seymour et al., 2019; Papin et al., 2021; Zador et al., 2019). However, these methods present significant drawbacks in the ICU context. Some of these are the need to include prior knowledge in the clustering models, accounting for sparse data and adapting to incomplete data.

In this chapter, we explore the use of non-parametric Bayesian community detection to identify data-driven patient profiles, aiming to mitigate the limitations previously mentioned in existing methods. We start by using widely used machine-learning approaches to identify multimorbidity profiles using the Elixhauser and Charlson indices. Then we introduce our approach to the same task from a community detection perspective. We end with an extensive comparison of results and discuss and conclude on the benefits and drawbacks of our proposed approach.

4.1 Clustering to manage the heterogeneity of patients

Clustering is widely used in the ICU to manage patient heterogeneity by identifying groups with similar clinical profiles. Extensive research has applied clustering techniques to identify patient groups based on multimorbidity profiles. Although clustering results vary due to the diversity of approaches and datasets, certain commonalities be found and serve as a starting point for our discussion.

A first aspect to explore is the number of clusters. This is relevant to our work since it is common to assume this number as a hyperparameter for clustering in the multimorbidity domain (see Section 2.2). Survey papers report that the typical number of clusters ranges between 4 and 10 (Busija et al., 2019; Prados-Torres et al., 2014). This is consistent with our observations in literature (Zador et al., 2019) and our results for traditional and our non-parametric clustering approach (see Sections 4.2.2 and 4.2.4).

Another element to note is the type of clusters in the literature. While these vary greatly, at least two types are consistently reported in literature (Busija et al., 2019; Prados-Torres et al., 2014): mental health cardiovascular and metabolic diseases. Interestingly, mental health diseases can be found in co-occurrence with substance abuse and kidney disease (Ballard, 1997; Salmon-Ceron et al., 2005). Cardiovascular clusters often include conditions such as hypertension, cardiovascular disease and diabetes (Wilson et al., 1999; Zador et al., 2019). Notable, as we show later in this chapter, we replicate these results both with the traditional and our proposed clustering approach.

However, as discussed in Section 2.2, challenges such as data quality and sparsity, the need for prior knowledge of the number of clusters, and model selection complicate the effective the effective use of clustering in the ICU. In Section 4.1.1, we review common ICU clustering methods used in literature, namely K-modes, Latent Class Analysis, and Hierarchical Agglomerative clustering. We then introduce hierarchical stochastic block modeling as an alternative for clustering in Section 4.1.2, discussing its benefits and highlighting its differences from traditional approaches.

4.1.1 Clustering techniques for healthcare.

In the following we revisit three of the most widely used clustering techniques in healthcare centroid based methods, LCA and Agglomerative clustering. While various versions of these methods exist, we focus our review on the most commonly used versions available in the literature. These methods are well-validated in the literature, allowing us to effectively use them as relevant benchmarks for our approach.

Centroid-based clustering

An intuitive approach to clustering is to rely on the observed characteristics of patients, defining the similarity between two patients as the degree to which they share similar characteristics. Partition-based methods, discussed in Section 2.2.1, formalize the idea of similarity as a distance between patients in the feature space (Murphy, 2012). Centroid-based clustering models (Xu and Tian, 2015; Ezugwu et al., 2022), a family of partition clustering algorithms, make use of the idea of distance to define clusters around centroids, a vector located in the centre of each cluster. Observations are assigned into a fixed number of clusters such that each observation is assigned to the cluster with its centroid closest to it. K-means, a widely used algorithm of this type, has been used for clustering in multiple settings such as in the classification of patients with bipolar disorder using demographic and clinical data (Fuente-Tomas et al., 2019) and to study groups of patients with similar morbidity co-occurrence and the potential impact of specifically adapted treatments (Seymour et al., 2019; Papin et al., 2021).

The reliance on the concept of distance makes it difficult to use K-means with categorical data (Murphy, 2012), common in the healthcare domain. An approach to deal with this is to modify the data to make it numerical and then apply K-means. For example, in a study of multimorbidity profiles, the prevalence of morbidities was considered for clustering instead of the raw information of the patient's existing comorbidities (Zador et al., 2019). A more principled solution is the application of K-modes (Huang, 1997), an application of the K-means paradigm to cluster categorical data. In this case the concept of distance is extended to include the number of matching variables between patients, and the centroids are computed using the mode for each categorical variable in each cluster instead of the mean (Huang, 1997, 1998). This approach has been used to identify profiles in kidney disease patients (Popoola et al., 2021) and in oncology patients with distinct symptoms profile (Papachristou et al., 2018).

These methods are used due to their conceptual simplicity and computational scalability (Raykov et al., 2016). However, they exhibit limitations particularly relevant in the healthcare domain. Firstly, they require prior knowledge in terms of the number of clusters, which is not always known. Secondly, they heavily rely on quality data which is not always available in the ICU. Thirdly, these methods assume independence between clinical features, which might not be true in the medical domain. Finally, distance metrics impose assumptions about the distribution of features, leading to potentially unexpected clusters (Raykov et al., 2016).

Latent Class Analysis

As discussed in 2.2.2, Latent Class Analysis (LCA) is one of the probabilistic approaches commonly used in the healthcare domain (Busija et al., 2019). This approach has been used to identify seven complex patient morbidity profiles with significantly different 1-year health care utilization and mortality (Grant et al., 2020). It has also been used on laboratory and demographic data to identify four multimorbidity profiles with different prevalence of sepsis and mortality (Zhang et al., 2018). Similarly, Zador et. al. identified six multimorbidity profiles (Zador et al., 2019).

Despite being a more principled approach and addressing some of the issues of centroid-based algorithms, LCA shares some of the same drawbacks: Firstly, it requires the use of heuristics to define the optimal number of clusters, which could lead to poor results (Nasserinejad et al., 2017; Busija et al., 2019). Secondly, LCA assumes conditional independence of the observed features given the clusters. This might not hold in the healthcare domain, as dependencies between morbidities might exist within clusters, consequently affecting the quality of the results (Busija et al., 2019). Finally, LCA uses unobserved variables to find clusters, making them hard to understand as they cannot be interpreted directly from the observed data (Moshkovitz et al., 2020).

Agglomerative hierarchical clustering

Hierarchical clustering, see Section 2.2.3, offers an alternative to flat clustering techniques, presenting a nested structure of clusters. This method relies on similarity metrics, such as Jaccard similarity coefficient (see Section 2.2.4), to group patients. Due to its frequent use in the healthcare domain, we focus on the agglomerative hierarchical clustering in this Chapter. This approach starts by assigning each patient to their own cluster and then iteratively merges similar clusters based on their similarity. This process continues until there is only one cluster, leaving a map of cluster models at each merging step which can be visualized as a dendrogram.

Unlike the previous approaches, agglomerative clustering does not require a predefined number of clusters. Moreover, the clustering outcome does not depend on initialization parameters, facilitating its reproducibility. Additionally, the hierarchical view of the clusters provides insights into the relationship between clusters and individual observations. Leveraging these advantages, this approach has been applied in various medical research contexts, including the identification of clinical profiles related to the onset of sepsis (Lvovschi et al., 2011) and in the identification of multimorbidity

profiles correlated to in-hospital mortality (Teh et al., 2018).

Despite its benefits, the agglomerative hierarchical approach also comes with drawbacks. Firstly, while it does not require prior information on the number of clusters, it lacks a principled method for determining the optimal number of clusters or hierarchical levels (Teh et al., 2018). Secondly, similar to the centroid-based methods discussed earlier, it assumes independence between features, a strong assumption in the healthcare domain (Busija et al., 2019). Finally, it is as robust to sparse data as the similarity function used, which means that it is prone to bias results in favour of dense features similar to centroid-based flat clustering approaches.

Summary of existing clustering approaches

As we have described, existing clustering approaches present various drawbacks hindering both their application and accuracy in the healthcare context. An added disadvantage of the methods presented so far is the lack of explainability derived from possible non-intuitive behaviours of the algorithms (e.g. K-modes) or due to the indirect relationship between the clusters and the observed data (e.g. in LCA). In the next section, we introduce the use of the stochastic block modelling for patient clustering in the healthcare domain. This method originates in network science, where it has been demonstrated as an alternative to clustering (Peixoto, 2019).

4.1.2 Community detection in networks for patient clustering

An alternative approach for modelling patients and their clinical features is to organize them as a network, where patients and features become nodes, connected by links representing their relationships (Restocchi et al., 2022). This network representation enables the direct encoding of complex relationships between patients and their associated variables, making it particularly suitable to model high level dependency between clinical variables, such as comorbidities, and patients (Barabási et al., 2011).

On top of this network representation, we make use of community detection, the analogue of clustering in network science (Javed et al., 2018), to identify underlying hierarchical structures in features and patients. Specifically, we employ the hierarchical stochastic block model (hSBM) (Peixoto, 2014b), a non-parametric approach that delivers a hierarchical clustering model capable of unveiling detailed clusters, as detailed in Chapter 2.3.

The hSBM has been shown as a good alternative for existing clustering models

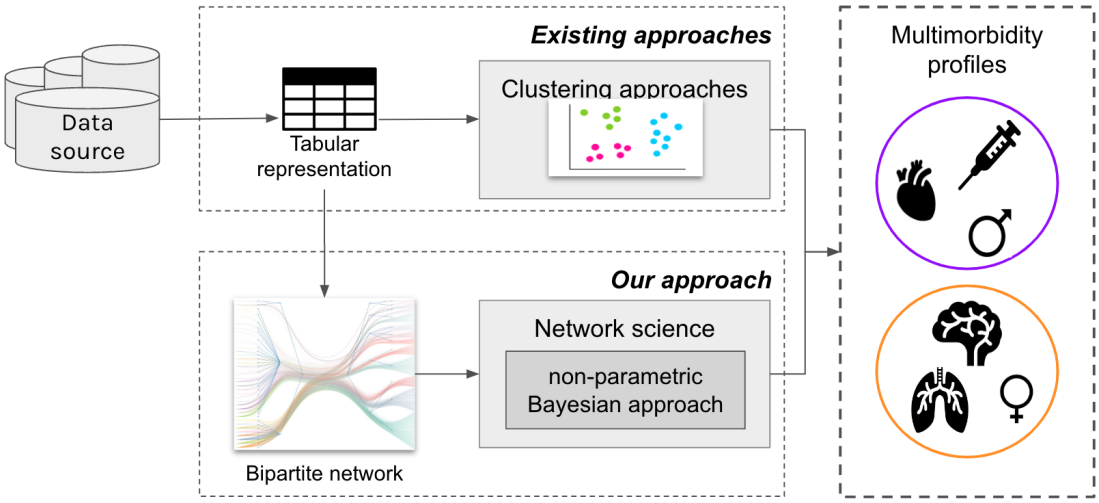

Figure 4.1: In contrast to existing approaches we propose a bi-partite representation of patients and comorbidities and the use of a non-parametric Bayesian approach for hierarchical community detection to uncover multimorbidity profiles.

(Peixoto, 2014b, 2017; Gerlach et al., 2018) and we believe that its advantages can be carried over into the healthcare domain. Firstly, the number of clusters is automatically inferred from the data. Secondly, the hSBM guarantees that the model cannot overfit or find structure where there is none (Peixoto, 2014b, 2017). Third, the hSBM relaxes all assumptions related to the distribution of features by using non-informative priors. Finally, its hierarchical approach does not struggle to find communities that might be too small for other approaches. Finally, hSBM automatically infers a hierarchical cluster structure that can facilitate the visualization of the solutions, similar to agglomerative approaches (Vellido, 2020b).

In the rest of this chapter, we examine the application of hSBM to the discovery of multimorbidity profiles for patients in intensive care units (ICU). We then compare these results with three other traditional clustering techniques.

4.2 Experimental setting

In this section we present the experiments conducted to assess the effectiveness of our approach in identifying multimorbidity profiles. Figure 4.1 outlines our experimental pipeline, presenting the differences between current approaches and ours. These experiments utilize the cohort of patients introduced in Section 3.3.

We begin by exploring the current approach for the identification of multimorbidity profiles. We explore the use of K-modes, LCA and Agglomerative Hierarchical

clustering as these have been methodologically validated in the medical literature and can serve as benchmarks for evaluating hSBM's performance. We start by briefly reintroducing our study cohort in Section 4.2.1. Then, in Section 4.2.2 we summarize the extensive work required to train and select optimal models for K-modes and LCA, and in Section 4.2.3 we do the same for the hierarchical agglomerative model. Finally, in Section 4.2.4 we employ our approach for the identification of multimorbidity profiles.

4.2.1 Patient cohort for the multimorbidity profile study

We perform our experiments using the cohort presented in detail in Section 3.3. Our data includes both demographic, admission type and multimorbidity conditions. The combination of these characteristics is common in the literature as it has been noted that the relashionship between demographics and multimorbidites (Zador et al., 2019). For example, the number of morbidities is relevant in the risk of mortality (Elixhauser et al., 1998; Charlson et al., 1987), but at the same time the number of comorbidities increases with age (Barnett et al., 2012a). It has also been reported the importance of admission type in mortality and sepsis (Zador et al., 2019). By combining all of these features we aim to explore the relevance of each in the clustering of patient and create a dataset that allows us to make a fare comparison with the relevant literature.

As a reminder, this cohort includes only first ICU admissions for patients over 16 years old to avoid considering a patients multiple times. For each patient, we consider their age, sex, admission type (elective, non-elective) and comorbidities (Charlson and Elixhauser). As presented in Section 3.1, following relevant literature the age of patients is discretized into the following ranges: 16-24, 25-44, 45-64, 65-84, and over 85. The final study cohort includes 38,417 patients. Using this cohort we created two different datasets, one including the Elixhauser comorbidity index and another with the Charlson comorbidity index. We use both comorbidity indices in our experiments to evaluate our approach's advantages across differences in the datasets.

4.2.2 K-modes and LCA to identify multimorbidity profiles

In this section we introduce the multimorbidity profiles identified using K-modes and LCA in our ICU dataset. The choice of these clustering techniques reflects their relevance in the ICU context discussed in section 4.1. The final morbidity profiles discovered are presented in Figures 4.12 and 4.10, summarized as heatmaps detailing the prevalence of feature in each cluster relative to its prevalence in the sample. In the

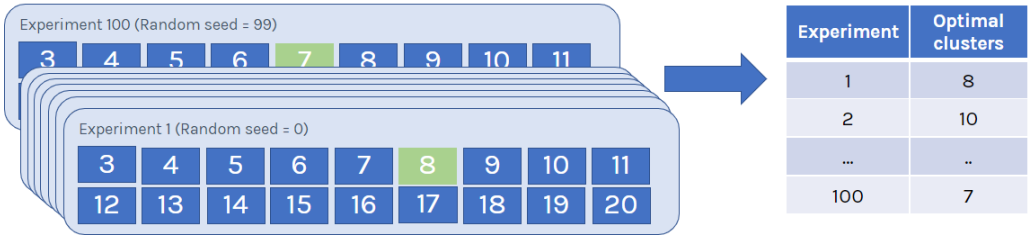

Figure 4.2: Model Selection: We conducted 100 experiments for each clustering method, each one including models with 3 to 20 clusters (blue boxes). Within each experiment, we determined the optimal number of clusters (green box) using the elbow method. This process yields a table displaying the optimal cluster numbers per experiment (on the right).

following, we describe the process of model selection and validation.

4.2.2.1 Model selection

To ensure the robustness of our models, we examine a broad spectrum of potential solutions in our selection process. To this aim, we run and examine 100 individual experiments for each model applied to the datasets with Elixhauser and Charlson comorbidity indices. Each experiment consists of building models ranging between 3 and 20 clusters (light blue boxes in Figure 4.2). The clustering range is aligned with relevant clinical literature (Åkerlund et al., 2022; Zador et al., 2019) and aims to strike a balance between having a minimum number of clusters that is significant and not as many as to create an unmanageable and over-fitted model.

We begin by determining the optimal number of clusters for each experiment (green box in Figure 4.2). To this purpose, we apply the elbow method (see Section 2.2.4) in each experiment using the number of clusters versus quality metric plot. As metric we use the Bayesian Information Criteria for LCA, for K-modes we use the distortion, the average distance of each observation to the centroid of the cluster it has been assigned to as defined in equation 2.4). We obtain 100 candidate models with different numbers of clusters, illustrated in the table in Figure 4.2. We chose the most frequent of the optimal clusters in this table as the best number of clusters for the model. As presented in Figure 4.3, the chosen number of clusters for both K-modes and LCA using the Elixhauser dataset is 8, and 8 and 7 for the K-modes and LCA respectively with the Charlson dataset.

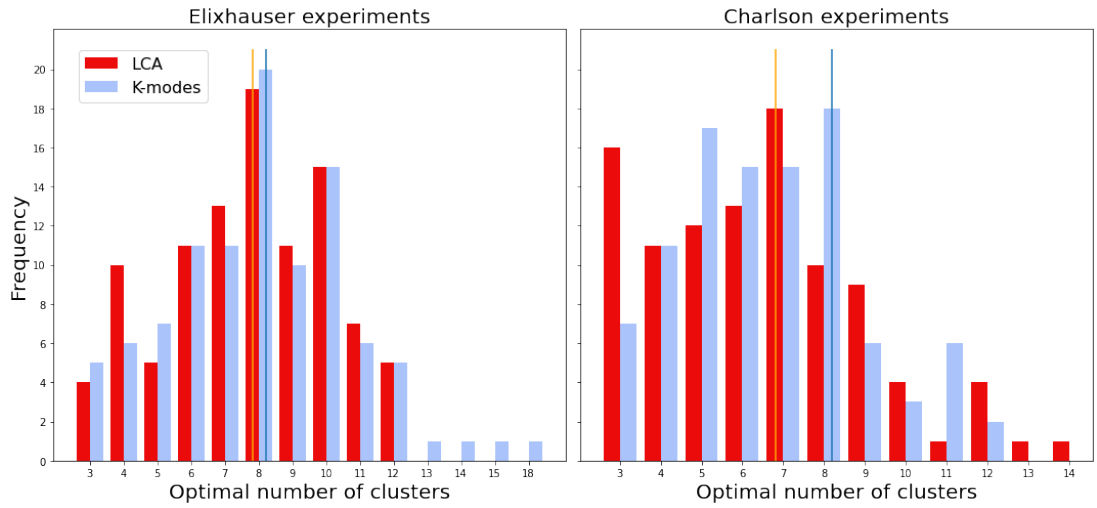

Figure 4.3: Frequency of optimal cluster numbers found using the elbow method in LCA (red) and K-modes (blue) experiments with Elixhauser (right) and Charlson (left) indices. The mode for LCA and K-modes experiments is denoted by the orange and blue vertical lines, respectively, representing our chosen optimal cluster count.

We finally select the model with the best performance metric among the candidates having the optimal number of clusters. The final models we choose following this methodology are presented in Figures 4.12 and 4.10 for the Elixhauser and Charlson datasets respectively. In the next section, we evaluate the stability of these clusters to validate the soundness of our selection.

4.2.2.2 Models stability

By scanning through various potential models and following the principle of entropy minimization we come to the most probable number of clusters identifiable from the data. Despite our approach providing an optimal model across multiple criteria, the question of stochasticity is still present. Is it plausible that our optimal solution is fortuitous, attributable to specific initial conditions? Might the lowest entropy be achieved with two or more vastly distinct partitions?

To address this issue we use the Mutual Information (MI) metric (see Section 2.2.4) to study the stability of the clustering models identified in the previous section. MI gauges the concurrence in membership allocation between two clustering models. This metric quantifies the similarity of two partitions, where a value of 1 signifies identical patient groupings and 0 no shared clusters. We use MI to evaluate the stability of the optimal models for both K-modes and LCA by comparing the top 10 models.

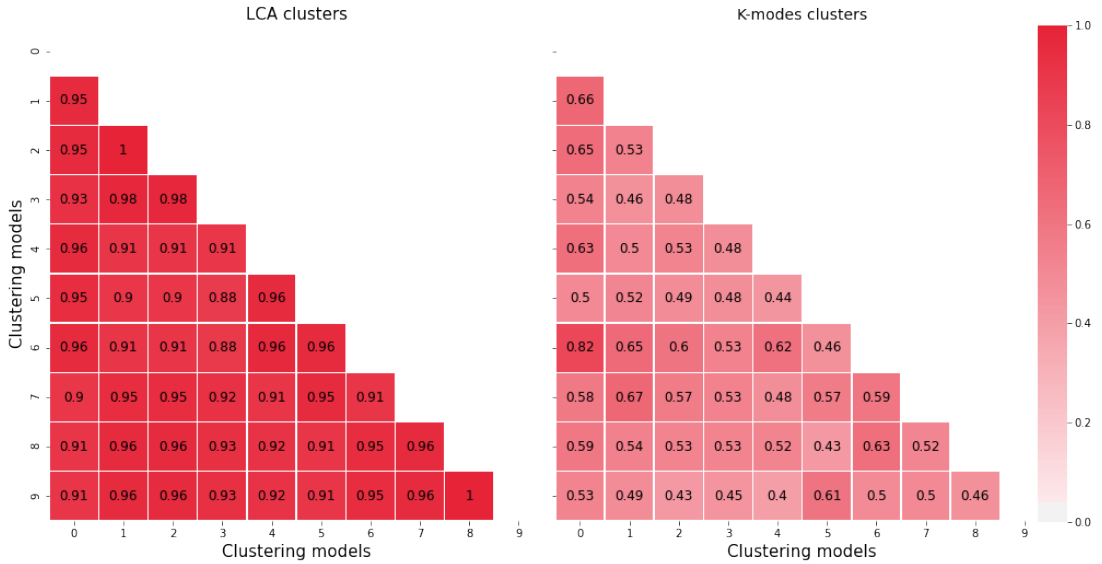

Figure 4.4: Mutual Information between the Top 10 Elixhauser models generated by LCA (left) and K-modes (right). Models are labelled from 0 to 9 based on entropy, with 0 being the lowest entropy and 9 the highest. Notably, MI is considerably greater in LCA partitions compared to K-modes partitions. The optimal model for each method (model 0) maintains the highest MI with other models, emphasizing its consistency across all models.

Figure 4.4 presents a comparison of MI between the top 10 multimorbidity profiles derived from Elixhauser data using LCA and K-modes. Interestingly, the LCA models exhibit higher MI values amongst themselves compared to the K-modes models. This suggests that LCA consistently groups patients in a similar manner, showing more stability compared to K-modes, where clustering vary more. The same can be observed for the Charlson dataset (see Appendix B.1), reinforcing this idea.

Regarding the stability of the selected models (named model 0 in both cases), it is worth noting that they demonstrate a high level of uniformity with other examined models. In the LCA scenario, the minimum MI reaches 0.9, while in the case of K-modes, it stands at 0.5. Similar trends are observed when considering the clusters based on the Charlson index, as outlined in Figure B.1 of Appendix B. With these results in mind, we proceed with our study utilizing the models presented in Figures 4.12 and 4.10. These models not only show low entropy but also capture a cluster structure that aligns with that of the other leading models.

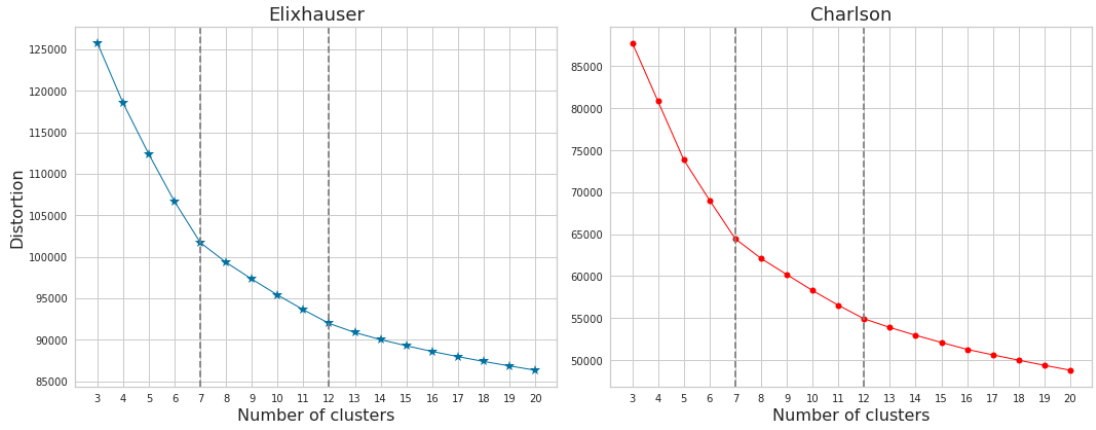

Figure 4.5: Elbow calculation for the agglomerative clustering using the Elixhauser comorbidity index (left) and Charlson (right). We identify two elbows, the first one at 7 clusters is computed in consideration of all cluster models. The second one at 12 clusters, considered only clusters between 8 and 20.

4.2.3 Agglomerative clustering to identify multimorbidity profiles

In the following, we present the outcomes of the agglomerative hierarchical clustering applied to the Elixhauser and Charlson multimorbidity datasets. This approach differs from the previous ones in that its outcome is deterministic given a similarity metric (see Section 4.1), thus there is no need to run multiple experiments for model selection.

We start by considering two metrics to measure similarity between patients. First, we employ the Jaccard coefficient (see Section 2.5.1), well-suited to our categorical dataset (Lvovschi et al., 2011). We also include the squared root of the difference between patients' features as an alternative distance used in the healthcare domain for one-hot encoded data (Cui et al., 2018). Using these metrics, we compute the distance matrix for each patient-patient pair. Based on this matrix, we build models with clusters ranging from 3 to 20 using an average linking strategy for cluster merging (see Section 2.2.3). These models are compared using the elbow method using the distortion score as a quality metric (see Section 2.5.1). We move forward using the patient distance as a merging metric since it yields the lowest distortion in all experiments.

For model selection, we favour the use of the elbow method to keep consistency with prior experiments and to reduce the subjectivity in the selection process. The elbow is applied twice. Initially, we consider all models (ranging from 3 to 20 clusters), revealing a first elbow at 7 clusters for both datasets. Then, we apply the elbow considering only clusters from 8 to 20 clusters, leading to another elbow at 12 clusters.

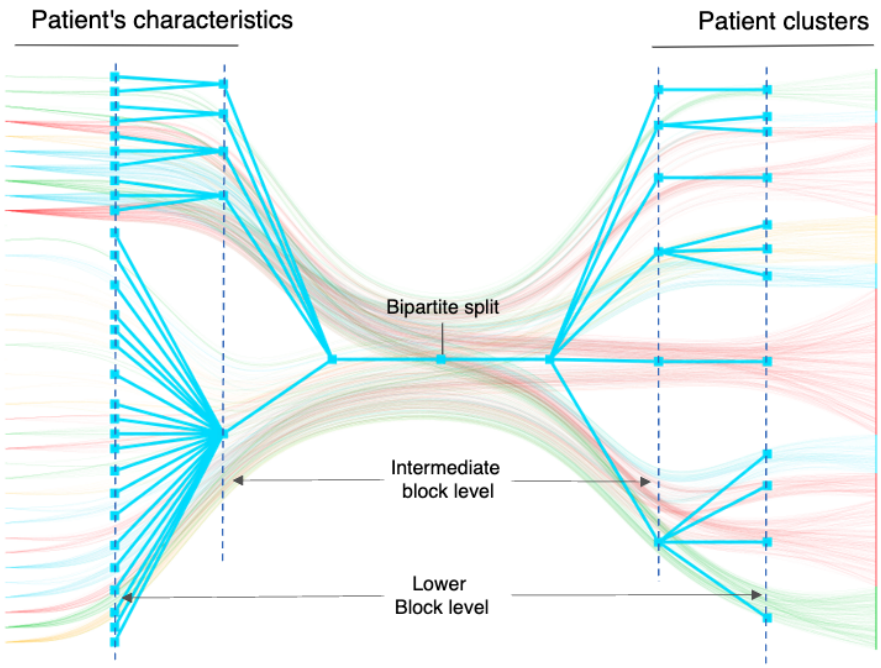

Figure 4.6: Bipartite network of patients and Elixauhser comorbidities and demographic features. Patients are displayed as nodes on the right and features on the left of the network. The tree illustrates the clustering structure inferred using the hSBM model, with 12 patient clusters at the lowest hierarchical level and 6 at the intermediate level.

We do not further apply the elbow as no significant elbow candidate is observed. As presented in Figure 4.5, this process results in a two-level hierarchical structure, with 7 and 12 clusters in both datasets. Notably, this structure aligns with the two-level hierarchical structure that the hSBM unveils (see Section 4.2.4). The final models are shown in Figures 4.12 and 4.10.

4.2.4 Community detection to identify multimorbidity profiles

In this section, we introduce our approach to identifying multimorbidity profiles in ICU patients using hierarchcial stochastic block modelling. We start by constructing a bipartite network connecting patients and their clinical features. This step represents the first difference with commonly used methods as the network-based representation of patients and features provides a more expressive means to encode their complex relationships. Next, we run the hSBM algorithm, discussed in Section 4.2.4, to infer the most probable hierarchical clustering structure using the Elixhauser and Charlson datasets. In the following we describe the process leading to our final models shown in figures 4.11 and 4.13.

We began by constructing a bipartite network to represent the relationship between patients and their characteristics. Figure 4.6 illustrates this network, with patients represented by nodes on the right and characteristics nodes (age group, sex, and Elixhauser comorbidity) on the left. Links connect each patient to their characteristics; for example, a male patient will be linked to "male" sex node. It is important to note that our network is unweighted as each link is considered equally important. Similarly, we build a bipartite network for the Charlson comorbidity index.

We then apply the hSBM to unveil cluster structures using a two-step methodology to avoid local and sub optimal solution (Peixoto, 2021). Firstly, we employ an agglomerative multilevel Markov Chain Monte Carlo (MCMC) algorithm (Peixoto, 2014a), which starts by partitioning all nodes in distinct clusters and then, at each step, proposes moving nodes between clusters. These moves are accepted with a probability based on their resulting reduction in entropy (MDL in our case), or entropy gain. Due to its stochastic nature, this algorithm cannot ensure the best partition. To mitigate this, we ran the algorithm 100 times. Each new model is compared to the best previous one by computing the ratio of their MLD scores. This posterior odds ratio quantifies how much more likely is the new model to fit the old one.

However, there is still a small chance that the algorithm converged to a local minimum. To address this, we further refine our solution using the merge-split MCMC algorithm (Peixoto, 2020), which addresses this issue by proposing to move groups of nodes instead of single ones. This allows for a more exhaustive scan of the solution space of group assignments. We run this 10,000 times, in batches of 10, to ensure that no further significant improvement is possible. After roughly 200 run batches the improvement in the posterior likelihood stabilizes (Figure 4.7), suggesting that the clusters obtained are either optimal or close to the optimal.

Figure 4.6 shows the final clusters. The tree-like structure overlaid on the network represents the hierarchical clustering structure found using the described methodology. Starting from the "Bipartite split," the right branches depict patient clusters and their hierarchy. At the lower block level, there are 12 patient clusters, represented by the 12 leaves on the right. These clusters are further organized into 6 at the intermediate-level clusters. Similarly, the left branches show the hierarchical structure of patient characteristic clusters.

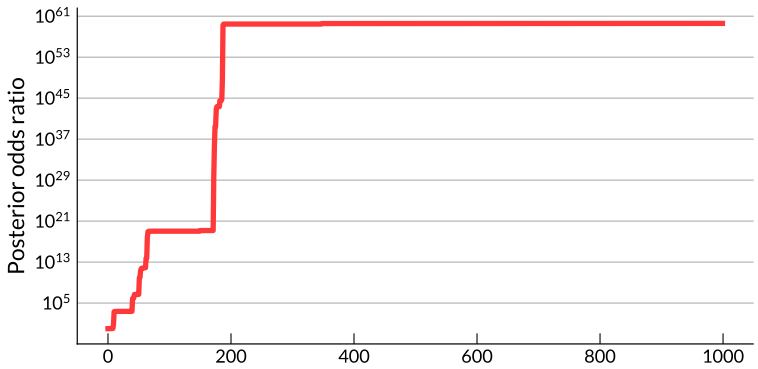

Figure 4.7: Posterior odds ratio obtained at every run of the merge-split MCMC. After around 200 batch runs the posterior odds ratio stabilises at a state whose partition is $\approx 10^{60}$ more likely than the one found initially by the agglomerative multilevel MCMC.

4.3 Clustering analysis

In this section, we conduct an in-depth analysis and comparison of our approach and the benchmark approaches, namely K-modes, LCA and hierarchical agglomerative clustering. We organize this section in four subsections to address the variety and intricacy of these models. We start with a quantitative analysis of our models, aiming to understand their overall quality and general behaviour. Then, we move into a detailed comparison of the clustering models.

Since a direct comparison between hSBM, hierarchical agglomerative models, and the flat partitions models of LCA and K-modes is hard to conduct, we break our analysis into two parts. We begin by evaluating the benchmark outcomes against each other and the intermediate cluster structure of the agglomerative and hSBM models. This comparison is possible due to the fact that hSBM and the agglomerative clustering models identify a comparable number of clusters to the flat methods using the Elixhauser and Charlson morbidity indices. Then, we examine the lower hierarchical level of the hSBM comparing it mainly with the lower hierarchical level of the agglomerative clustering model.

4.3.1 Quantitative analysis of the clustering models

We start by comparing the clustering models presented in the previous section using a quantitative approach. A common criterion to quantify the quality of cluster partitions is to examine the level to which elements within clusters are similar and at the same time different to those in other clusters. In this section, we use the Caliński-Harabasz

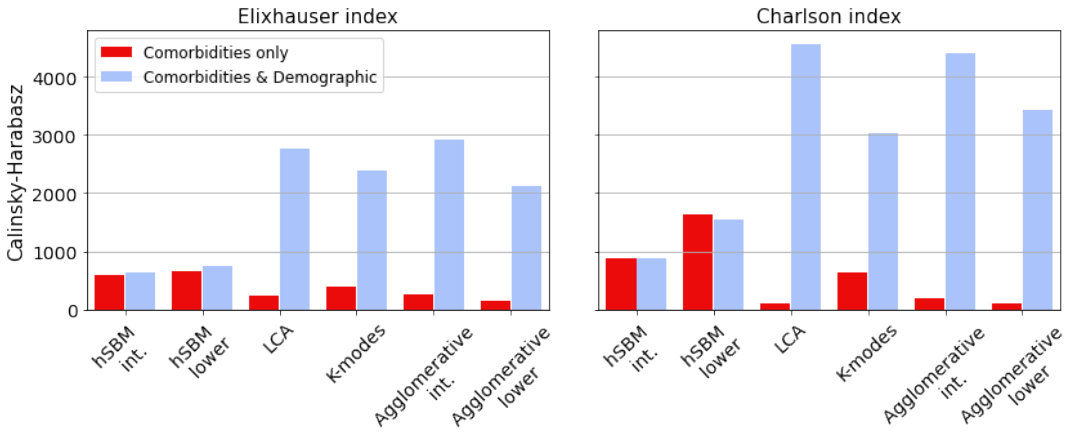

Figure 4.8: Calinski-Harabazs score comparison among the different clustering approaches for the Elixhauser (on the right) and Charlson (on the left) comorbidity indices. Notably, LCA, K-modes and the agglomerative clustering are greatly impacted by the presence of the demographic and admission type information. In contrast, hSBM performance remains consistent, suggesting its ability to incorporate the relevance of the comorbidity in the final clustering model.

(CH) score (Caliński and Harabasz, 1974b) score (Section 2.5.1) to evaluate the K-modes, LCA, hierarchical agglomerative and hSBM models previously introduced. In Figure 4.8, we show the CH score for the clustering models derived using the Elixhauser and Charlson comorbidity indices.

On top of examining the overall quality of the different models, we conduct two experiments for each dataset to assess the impact of morbidities and demographic data separately on the clustering algorithms. As noted in Section 3.3 there is a noticeable difference in the density of morbidity features compared to demographic ones. While patients generally present relatively few comorbidities, each patient will always have records for age, sex and admission type. For instance, on average, patients exhibit 3 comorbidities out of the 30 possible Elixhauser comorbidities (only 10%), but every patient has a record for age, admission type and sex. This could potentially bias the clustering algorithms to overemphasize the denser features (i.e. demographics and admission type) at the expense of the sparser comorbidity features.

In the first experiment, the CH score was calculated considering all the features used to infer the clustering models (comorbidities, demographics, and admission type), presented in blue in Figure 4.8. For the second one, we compute the CH score considering only comorbidities, with results shown in red in the same figure. Results in

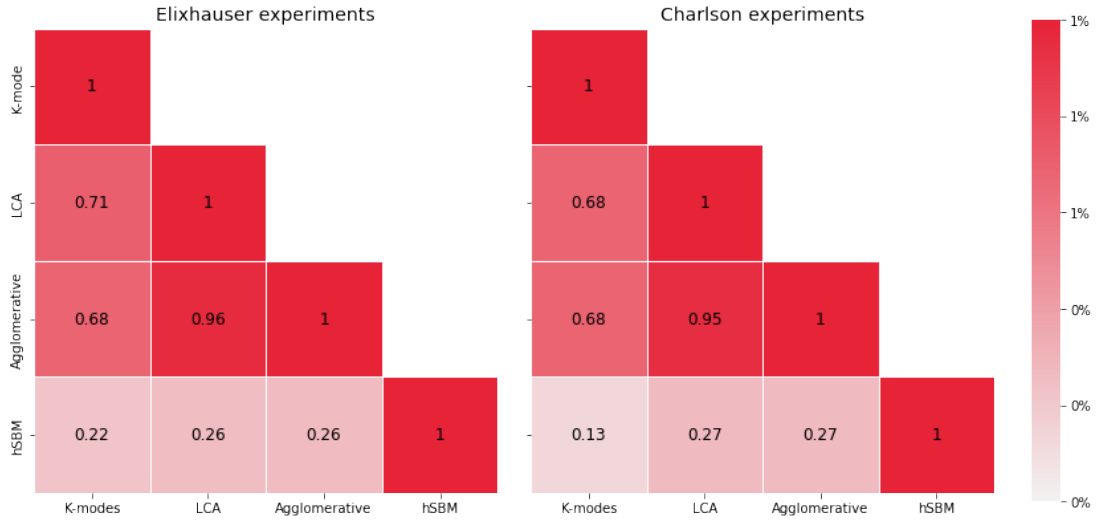

Figure 4.9: Mutual Information between the final clustering models for K-modes, LCA and agglomerative and hSBM at the lowest hierarchical level. Noticeably, hSBM consistently shows lower MI compared to the benchmarks. This supports the idea that hSBM can identify different clusters than those found by traditional methods.

Figure 4.8, show the impact of demographic and admission type features on the LCA, K-modes and the agglomerative models, greater than on the hSBM. The impact is such that these models go from presenting the best CH score when considering all the features, to being the worst when considering only comorbidities. Maybe more surprising is that for the hSBM, the CH score virtually remains the same across experiments, pointing to the robustness of the hSBM clusters to variations in the denser features.

The previous observations suggest that while the LCA, K-modes and agglomerative hierarchical models rely on the demographics and admission type to infer the similarity between patients, the hSBM can identify relevant multimorbidity patterns connecting patients even though these are sparse. In the upcoming sections, we will delve into this further as we analyze the prevalence of features within each cluster model. This analysis will reveal that demographic features predominantly determine the clusters generated by LCA, K-modes and the agglomerative approaches, whereas hSBM strikes a balance across all features.

4.3.2 High level comparison of multimorbidity clusters

We start our analysis by comparing the LCA and K-modes clustering models to the highest hierarchical level of the agglomerative and hSBM models. We first present

a comparison between the benchmark methods to then incorporate our findings with the hSBM model. To facilitate the analysis, we assign a suffix to each cluster in the text, indicating whether it belongs to the Elixhauser or Charlson cluster. For instance, cluster $L1_e$ denotes Elixhauser LCA cluster number 1, while $L1_c$ denotes Charlson LCA cluster number 1.

4.3.2.1 Benchmark clustering model comparison

Our analysis of clusters brings to light the pronounced role of demographics and admission type across the different benchmark methods and indices. As we describe below, despite their methodological difference LCA, K-modes and hierarchical agglomerative clustering methods all yield very similar clustering structures. The latter shows clear groupings based on age, sex and admission type, but is not as clear when examining their comorbidity composition.

A notable example is the relevance of admission type. Elixhauser clusters $L7_e$, $L8_e$, and $K8_e$ and Charlson clusters $L3_c$ and $K8_c$, in Figure 4.12 and 4.10 respectively, are characterized by a very high prevalence of elective patients. This pattern persists in hierarchical clustering with clusters HC-A_e and HC-A_c as shown in Figures 4.12 and 4.10. Interestingly, Elixhauser clusters $L8_e$ (6.31% of the sample) and $L7_e$ (9.39 % of the sample) are gender specialisation for elective patients, with the former including mostly female patients and the latter male.

Another similarity across models and indices is the significance of age and gender. Aligned with the literature (Zador et al., 2019; Papin et al., 2021), all models show groups of younger male patients with AIDS and varying prevalence of liver disease. These clusters are Elixhauser $L4_e$ and $K5_e$ (13.19% and 10.19% of the sample respectively) and agglomerative clusters HC-E_e (13.12% of the sample). A similar trend is observed in Charlson clusters HC-G_c and $L7_c$ (13.18% and 7.94% of the sample respectively), and to some extent in $K1_c$. These clusters show slightly lower rates of sepsis and mortality, likely due to the younger age of the patients. Notably, Elixhauser clusters $L2_e$, $K2_e$, and HC-C_e (16.80%, 20.63%, and 16.75% of the sample, respectively) and Charlson clusters HC-B_c and $K6_c$ (16.82% and 25.28% of the sample) cluster male patients with a similar morbidity profile but higher prevalence of liver disease as the previously discussed clusters and of older age. This suggests these clusters reflect a later stage of the younger patients with AIDS and liver disease.

Specific clusters for male and female patients between 65 and 84 years old are also unveiled by all the benchmark algorithms and across both comorbidity indices. These

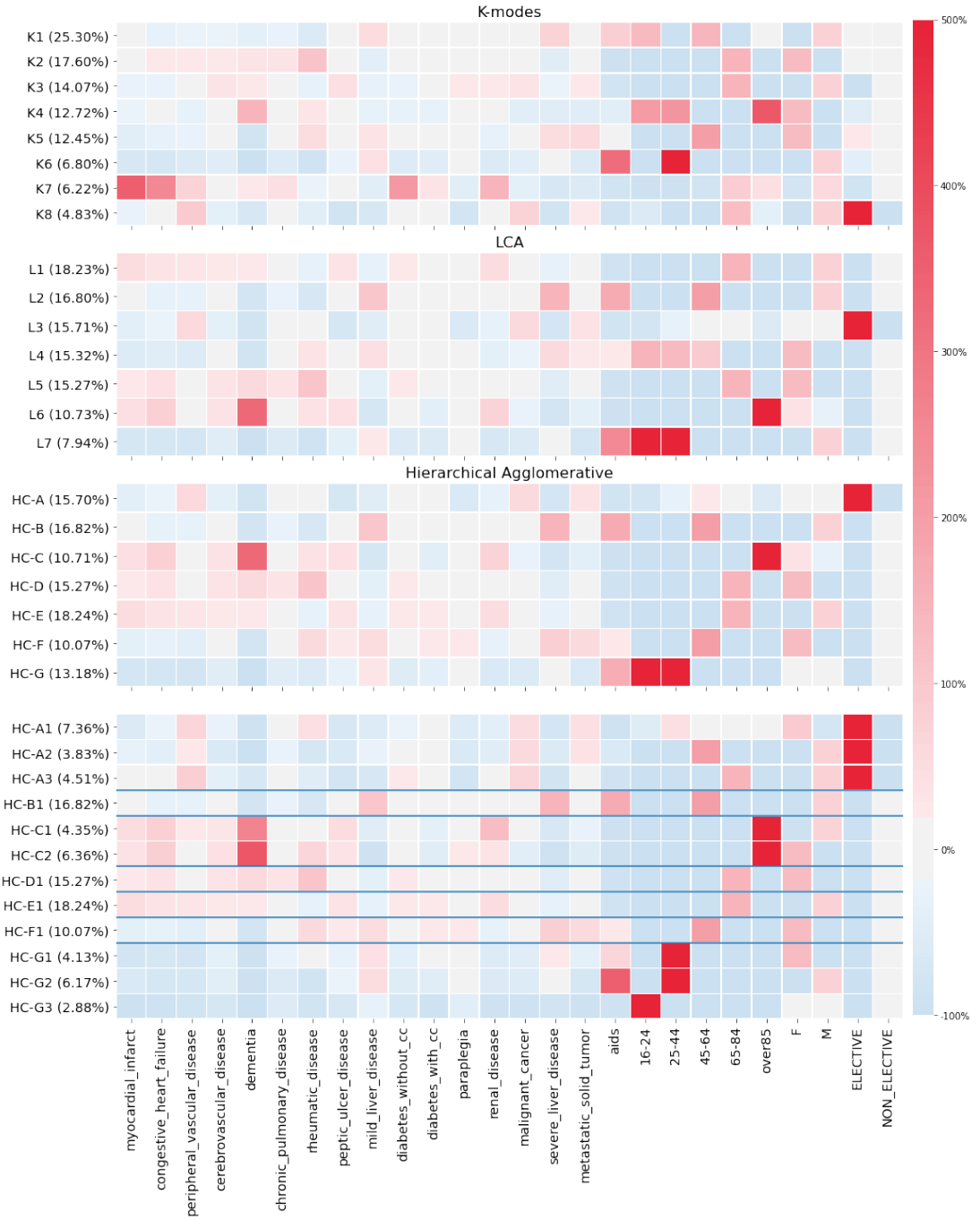

Figure 4.10: Best Charlson patient clusters for K-modes, LCA and hierarchical agglomerative clustering approaches. The heatmaps show the relative difference in features prevalence between each profile and the total populations, with red indicates higher prevalence and blue lower. A value of 0% for a cluster represents equal feature prevalence to the population, while 500% indicates a prevalence five times higher than the population. Similarly to the Elixhauser case in Figure 4.12, demographics and admission type exhibit an important role in the definition of profiles. For instance, elderly patients or non-elective patients are the main characteristics of clusters across methods.

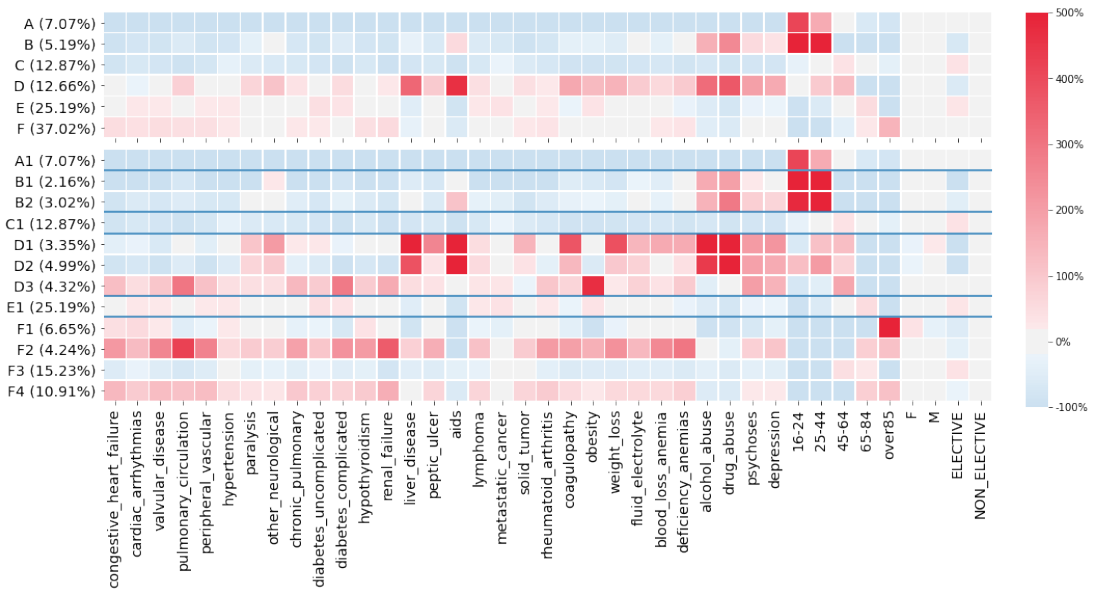

Figure 4.11: Composition of hSBM clusters for the Elixhauser dataset. At the top are the clusters from the intermediate hierarchical level, which divides into clusters at the low hierarchical level (indicated by lines between clusters). For instance, cluster B divides into B1 and B2. The heatmap displays the relative morbidity prevalence between each cluster and the entire dataset. A value of 500% represents a cluster prevalence 5 times higher than in the entire population.

clusters are Elixhauser clusters $K1_e$, $L1_e$ and HC-D_e (21.38%, 18.23% and 13.32% of the sample) grouping the male patients and $K4_e$, $L3_e$ and HC-F_e (11.66%, 15.27% and 15.32% of the population) doing the same for female patients. These clusters appear as well in the Charlson index. The more evident examples are clusters $K3_c$, $L1_c$ and HC-E_c (14.07%, 18.23% and 18.24% of the population) for male patients and $K2_c$, $L5_c$ and HC-D_c (17.60%, 15.27% of the population in both models) for female patients.

Clusters of female patients between 45 and 64 years old represent another similarity between algorithms and indices. This grouping appears in Elixhauser clusters HC-G_e and $L6_e$ (10.07% of the sample for both clusters) and K3 (17.63% of the sample) although in this last case, the cluster extends the age group including female patients under 64 years old. Similar clusters are observed when the Charlson index is considered despite the comorbidities being different. Charlson cluster HC-F_c (10.07% of the sample) is notably similar to Elixhauser clusters HC-G_e and $L6_e$ while Charlson cluster $L4_c$ and $K5_c$ (15.32% and 12.45% of the sample) are comparable to Elixhauser $K3_e$ as both include female patients under 64 years old.

While not as strongly as in the previous examples, the prevalence of patients over

85 years old play a role in the definition of clusters across different benchmark approaches and datasets. Starting with the Elixhauser index, clusters $K6_e$, $L5_e$ and $HC\text{-}B_e$ (7.92%, 10.73%, and 10.73% of the sample) represent this age group with high prevalence of congestive heart failure, cardiac arrhythmia, valvular disease and deficiency anaemia, hypothyroidism, and renal failure. Notably, this same grouping appears in Charlson clusters $L6_c$ and $HC\text{-}C_c$ (10.73% and 10.71% of the sample), but this time accompanied by high prevalence of dementia and lower, but still high, prevalence of congestive heart failure and renal disease. Interestingly, the effect of this age group is not that substantial when examining the Charlson K-modes partitions. In these cases, clusters include other patients in the age range of 16 to 64 and are specialized by gender. These clusters are $K5_c$ for female patients (12.45% of the sample) and $K7_c$ (6.22% of the sample) for male patients.

Perhaps the most notable difference between the benchmark methods is that K-modes identify a group of patients characterized by an elevated occurrence of uncomplicated diabetes and congestive heart failure that the rest of the models fail to unveil. These are the cluster $K7_e$ and $K7_c$ (5.72% and 6.22% of the sample). Interestingly, despite sharing a common core of significant comorbidities these clusters are quite different in terms of demographics. The Elixhauser cluster shows a high number of female patients between 65 and 84 years old and the Charlson one has a high prevalence of male patients over 65 years old.

Despite the last example, strong similarities exist between the K-modes, LCA and the agglomerative hierarchical approach at its higher level. These similarities underscore the influential role of demographics and admission type in the clustering. As discussed in Section 2.2, this can be explained due to the higher density of these features compared to multimorbidities. These findings align with our observations from Section 4.3.1, highlighting the heavy reliance of benchmarks on denser features, i.e. demographics and admission type. While this does not mean that the partitions found are incorrect, these methods might overlook important comorbidity patterns related to medical outcomes and that are influenced by demographic characteristics (Barnett et al., 2012b). In the following section, we show how our approach addresses this limitation, capturing the same information traditional methods extract while unveiling clusters based primarily on multimorbidities.

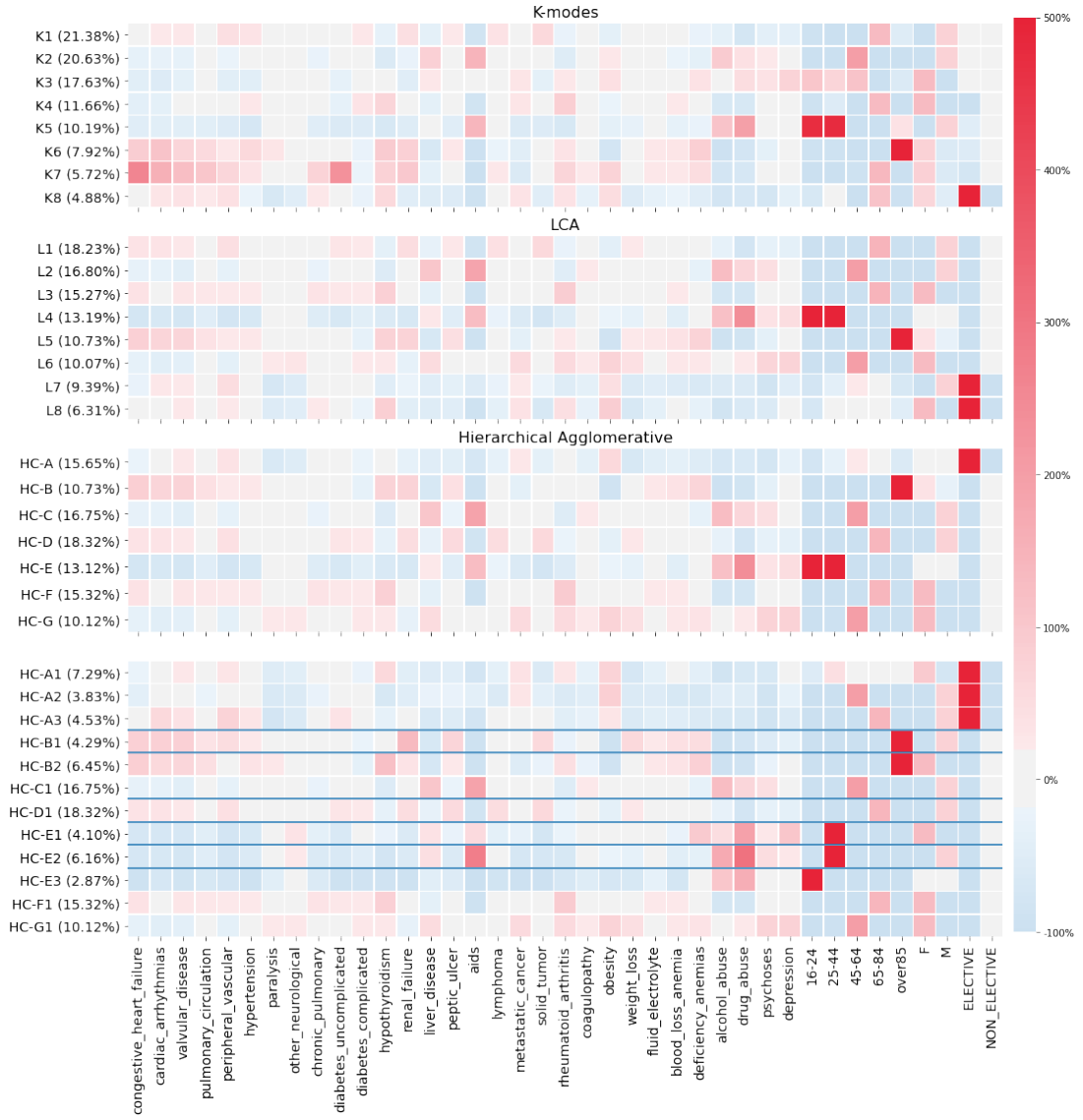

Figure 4.12: Best Elixhauser patient clusters for K-modes, LCA and hierarchical agglomerative clustering approaches. The heatmaps show the relative difference in feature prevalence between each profile and the total populations, with red indicating higher prevalence and blue lower. A value of 0% for a cluster represents equal feature prevalence to the population, while 500% indicates a prevalence five times higher than the population. Demographic factors and admission type significantly characterize profiles across methods. For example, patients over 85 years old are a primary characteristic for clusters in all presented methods.

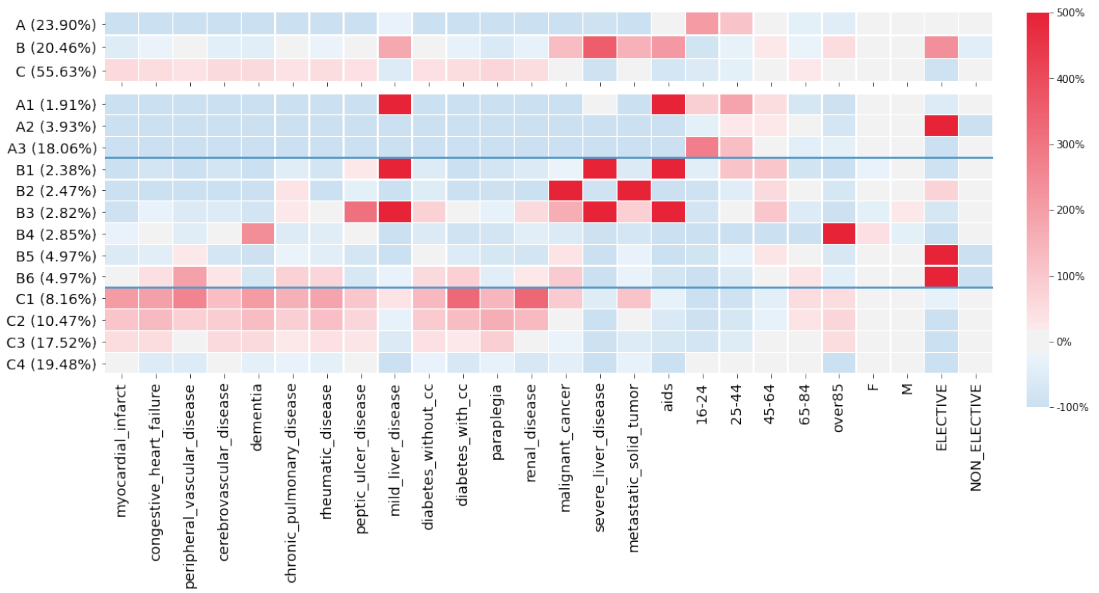

Figure 4.13: Composition of hSBM clusters for the Charlson dataset. At the top are clusters at the intermediate hierarchical level, dividing into clusters in the lower level as indicated by lines between clusters.

4.3.2.2 hSBM and benchmark clustering comparison

We start our analysis by looking at the level of agreement in patient cluster assignments between hSBM and benchmark methods. We use the Mutual Information score (MI), as discussed in Section 2.5.1, a higher MI indicates greater similarity in the clustering methods. Figure 4.9 shows the MI between hSBM and K-modes, LCA, and the Agglomerative methods at their lowest hierarchical level. Notably for both comorbidity indices, hSBM has low MI scores with all other methods, suggesting it groups patients different. On the contrary, the benchmark methods achieve higher MI scores between them, indicating a high level of agreement amongst them.

The above suggests that hSBM clusters patients differently than the benchmark methods. One explanation for this is that hSBM places less emphasis on demographic variables, as shown in Figure 4.8, and can include morbidities in the clustering process. This results in a model that simultaneously incorporates the effects of demographics and morbidities. Thus, our model identifies known clusters detected by traditional approaches and defined by demographic variables (e.g., clusters of patients over 85 years old) while also discovering new clusters. Our analysis begins by highlighting the similarities between hSBM and the benchmarks before emphasizing the novel information revealed by hSBM.

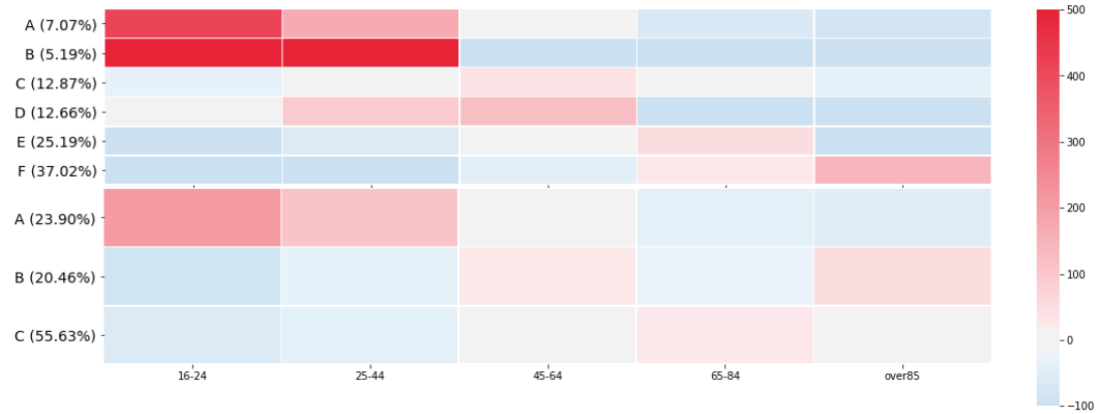

Figure 4.14: Age prevalence in hSBM clusters identified across morbidity indices. On top is the Elixhauser index and on the bottom is the Charlson. Notably, in both cases, younger patients tend to be grouped in few clusters.

Notably, all the approaches identify clusters strongly characterized by age across indices. For example, hSBM aggregates most patients in the age ranges 16-24 and 25-44 in Elixhauser clusters A_e and B_e and Charlson cluster A_c (see Figure 4.14). This pattern is also evident in the benchmark methods, particularly in Elixhauser clusters $L4_e$, $K5_e$, and HC-E_e and Charlson clusters $L7_c$, $K2_c$, $K5_c$, and HC-G_c. Another parallel lies in the treatment of older patients by the hSBM and benchmark methods. Clusters with a very high prevalence of patients between 65-84 years old can be found in the hSBM Elixhauser clusters E_e and Charlson C_c and benchmark clusters $L1_e$, $L3_e$, $K1_e$, $K4_e$, $K7_e$, $K8_e$, HC-D_e and HC-F_e and $L1_c$, $L5_c$, $K3_c$, $K4_c$, $K6_c$, $K8_c$, HG-D_c and HG-E_c. Similarly, clusters mainly characterised by patients over 85 years old can be seen in the hSBM Elixhauser cluster F_e and Charlson B_c, benchmark Elixhauser clusters $L5_e$, HC-B_e and Charlson clusters $L6_c$, $K7_c$, and HC-C_c. The consistency of these clusters across approaches is evidence their relevance in the ICU.

Beyond the mentioned similarities, various key differences set hSBM apart from our benchmark models. Notably, hSBM is able to consider all features for clustering, the interweaving of admission type with other morbidity features is one example. As a case in point, hSBM Elixhauser clusters B_e and D_e group patients with a lower prevalence of elective admissions (roughly 60% below average) combined with a high prevalence of alcohol and drug abuse. In contrast, the benchmarks group most elective admissions in Elixhauser clusters $L7_e$, $L8_e$, $K8_e$ and HC-A_e, offering no substantial insights into these patients' profiles. Similarly, Charlson hSBM cluster B_c merges patients with elective admissions (240% above average) and the highest prevalence of

liver disease and cancer. Benchmarks group these patients in clusters $L3_c$, $K6_c$, HC-A_c, with only marginally elevated cancer and peripheral vascular disease prevalence but lack the relevance of admission type.

Another difference lies in the significance of gender. While hSBM does not emphasize this feature, it remains relevant in the benchmarks. For instance, consider patients aged 65 to 84. Elixhauser clusters $L1_e$, $L3_e$, $K1_e$, $K4_e$ and HC-D_e and HC-F_e divide this group mainly by gender. A similar pattern emerges in the lower hierarchical level, where hSBM prioritizes the comorbidity profile over gender. This leads to Elixhauser clusters $E1_e$, $F1_e$, $F2_e$ and $F3_e$ and Charlson clusters $B6_c$, $C1_c$ and $C2_c$ with distinct prevalence of outcome. The same trend appears in the context of the Charlson index in clusters $L1_c$, $L5_c$, $K3_c$, $K4_c$ and HC-D_c and HC-E_c. In these clusters, despite gender specification, no significant insight is evident. In contrast, hSBM gives more relevance to comorbidity composition, yielding cluster C_c with the highest sepsis and mortality levels at the intermediate hierarchical level.

A final difference is the patterns of morbidity count in patients, a significant factor affecting patient outcomes (Charlson et al., 1987; Elixhauser et al., 1998; Busija et al., 2019). Figures 4.16 shows that benchmarks do not differ from each other in multimorbidity count, suggesting that this aspect is not captured (Further evidence for Charlson index in Figure B.6) shows similar trend with Charlson index). In contrast, our results reveal hSBM's capacity to group patients not only based on demographics and admission type but also by considering patients' individual morbidities count (excluding any other patient characteristic), as shown in Figure 4.15. Noticeably, this trend is constant across indices, with a clear distinction between clusters grouping patients having zero, few or multiple morbidities. This is clear in Elixhauser clusters D_e, E_e, F_e and Charlson clusters $B1_c$ to $B3_c$, $B6_c$ and $C1_c$ to $C3_c$ comprising patients with multiple morbidities (≥ 2), and Elixhauser clusters $B1_e$, $C1_e$ and Charlson clusters $A1_e$, $B4_e$, $B5_e$, and $C4_e$ with patients having only one comorbidity. Interestingly, we observe that the number of morbidities keeps a relationship with the type of morbidity. For example, patients in clusters with only one morbidity typically present the same type of morbidity. In Cluster $B1_e$, patients only present one psychosis, depression, alcohol or drug abuse or other neurological.

A final noteworthy finding of our approach is that none of the patients in clusters A_e, $A2_c$ or $A3_c$ have morbidities. These clusters represent younger patients with no long-term conditions that are completely undetected by the benchmark methods. Importantly, this category of patients is expected and of particular interest, but is un-

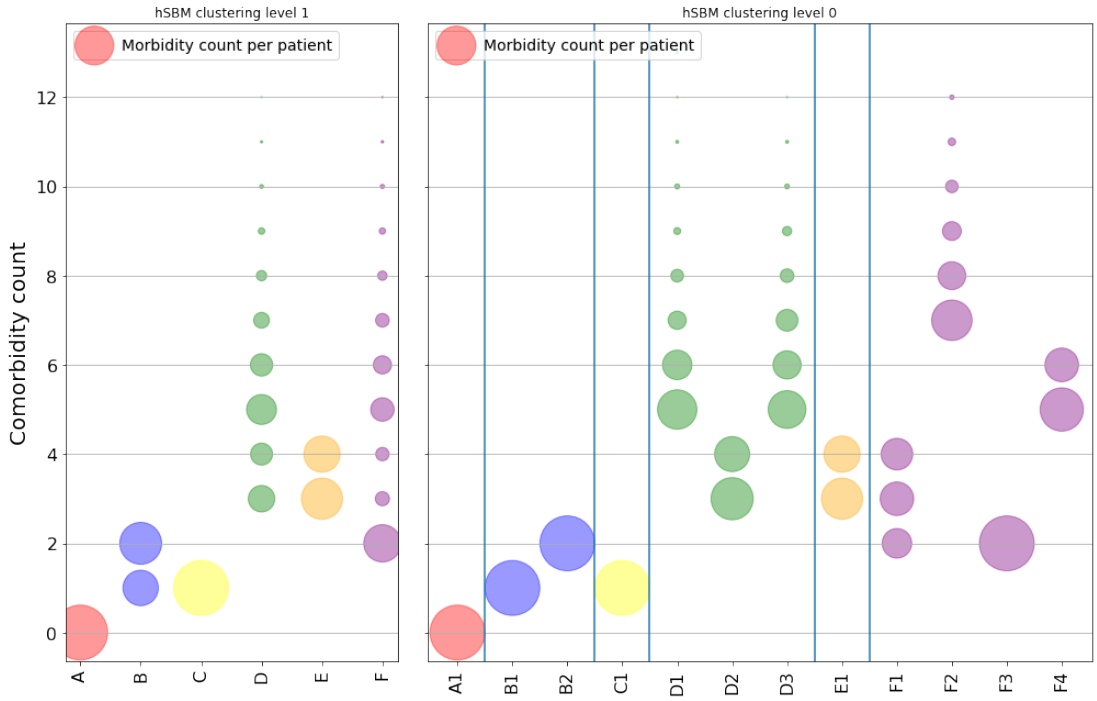

Figure 4.15: Count of Elixhauser morbidities per patient, i.e. the number of co-existing long-term disorders, at both the intermediate (left) and bottom (right) hierarchical levels. Circle size represents the number of patients with a specific count of morbidities. Surprisingly, we observe clusters, like A1, B1, C1, etc., where patients have only a few or no long-term illnesses. This indicates that the number of morbidities plays a crucial role in cluster formation. However, as the multimorbidity count increases, as seen in clusters D1, D3, and F2, the specific count appears less significant.

detectable by LCA, K-modes or the agglomerative models. This is also reflected in the much lower than average mortality and Charlson and Elixhauser comorbidity scores, shown in Figures 4.21 and 4.20.

The benefits of hSBM start to become evident at this level. Firstly, the method provides a more detailed description of the data and insights that the benchmark methods fail to capture. As alluded to previously, hSBM is able to combine sparse and dense features in the clustering assignment. Secondly, hSBM is able to capture non-explicit aspects of the dataset such as the multimorbidity count. Finally, hSBM is able to identify relevant groups that benchmarks fail to capture, such as clusters of patients without comorbidities. In the next section, we examine in detail the fine-grained results provided by hSBM.

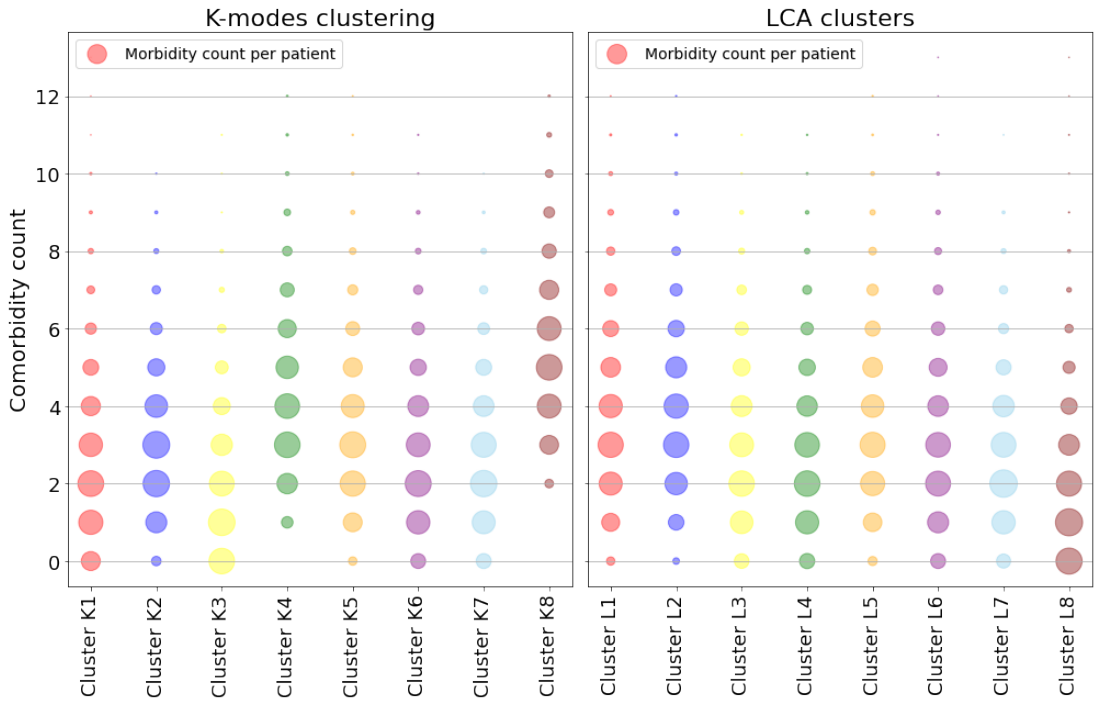

Figure 4.16: Number of Elixhauser-defined comorbidities per patient for LCA (right) and K-modes (left) clusters. Circle size represents the frequency of patients with a specific number of morbidities.

4.3.3 Comparison of fine-grained clusters

The previously mentioned hSBM clusters further divide into more detailed clusters at the lowest hierarchical level. Specifically, the six Elixhauser hSBM clusters break into 12 and the three Charlson ones split into 13. These clusters discriminate patients even further into more granular clusters, providing a more comprehensive understanding of patient groupings. Our analysis begins by comparing the hSBM to the agglomerative hierarchical model due to its shared hierarchical structure. We then move into a detailed analysis of hSBM clusters, focusing on their morbidity composition and count, the mortality and sepsis prevalence of the patients they group, and when possible in comparison with the benchmarks.

4.3.3.1 Hierarchical clustering comparison

Besides sharing a hierarchical structure, hSBM and agglomerative algorithms present dramatically different outcomes. Notably, the agglomerative hierarchical clustering model produces nearly identical clusters at the intermediate and lower hierarchical levels, despite differences in morbidity. In contrast, hSBM presents distinctly clustering

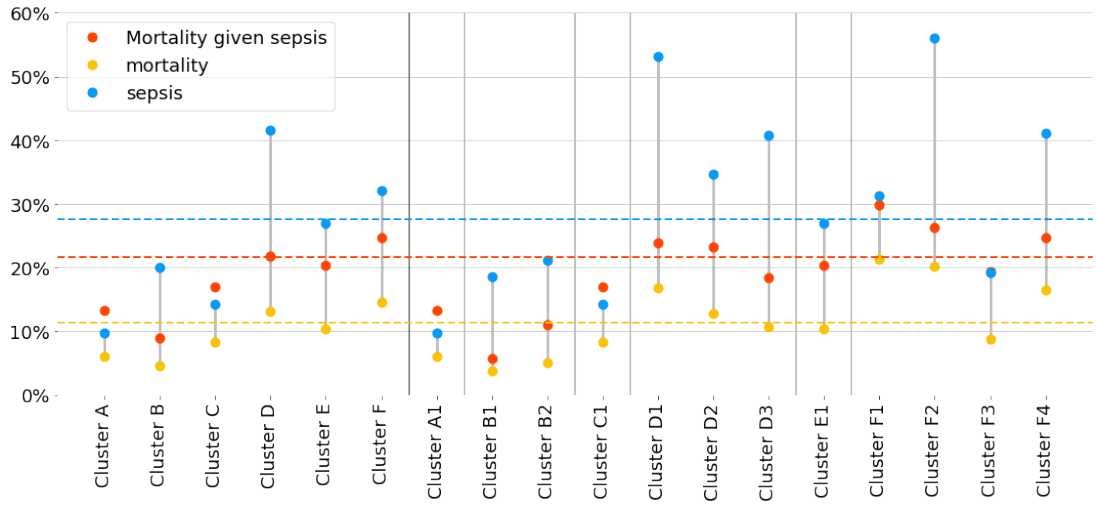

Figure 4.17: Heterogeneity in sepsis, mortality, and mortality-given-sepsis among Elixhauser hSBM clusters. The dotted lines show the average for mortality, sepsis and mortality given sepsis for the test cohort. Notably, at the finest hierarchical level, clusters like D1, F1, and F3 exhibit markedly higher sepsis and mortality rates than the subset average. Conversely, clusters such as those from A1 to C1 indicate a substantial prevalence of patients with low sepsis and mortality risk.

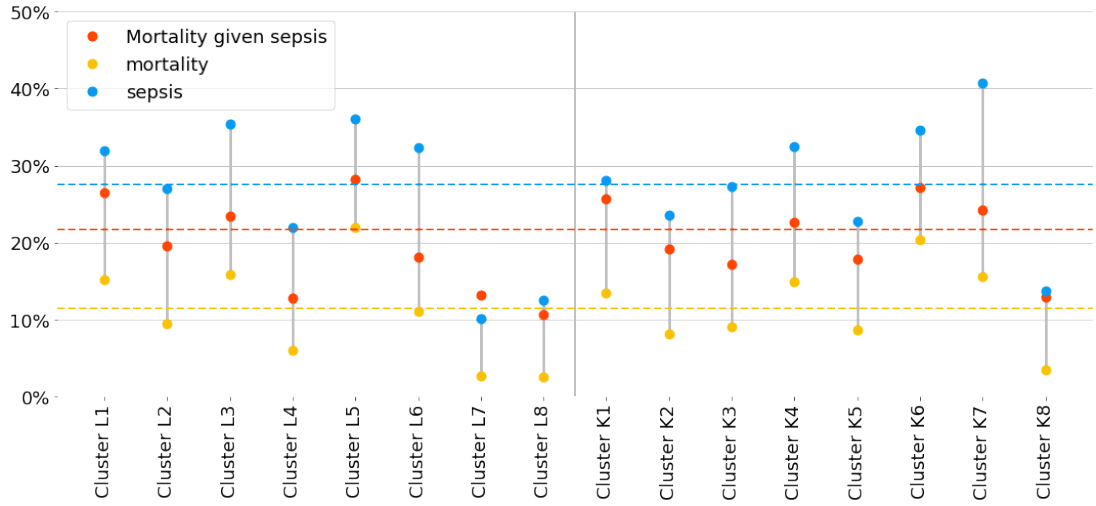

Figure 4.18: Sepsis, mortality, and mortality given sepsis across Elixhauser LCA (left) and K-modes (right) clusters. The dotted lines show the average for mortality, sepsis and mortality given sepsis for the test cohort. We notice greater stability in medical outcomes across clusters, except for clusters K2 and K6. This homogeneity suggests a limited association between these clusters and medical outcomes

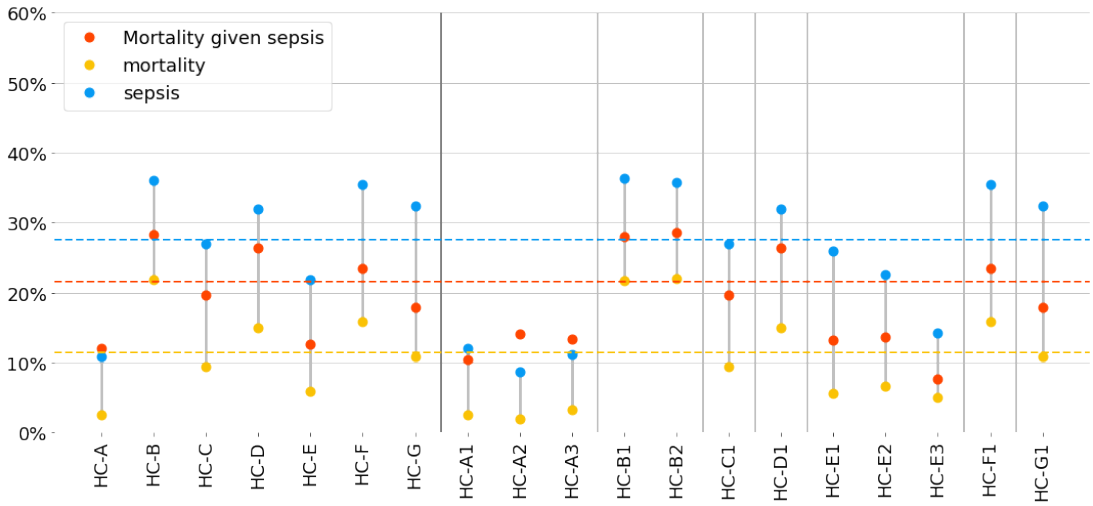

Figure 4.19: Sepsis, mortality, and mortality given sepsis across Elixhauser agglomerative clustering, 7 clusters (left) and 12 clusters (right). The dotted lines show the average for mortality, sepsis and mortality given sepsis for the test cohort. We notice greater stability in medical outcomes across clusters. This homogeneity suggests a limited association between these clusters and medical outcomes.

structures within the context of the Elixhauser and Charlson indices.

One of the most surprising findings is that agglomerative clusters can be mapped one-to-one between indices in terms of the demographics and admission type and the percentage of the population they encompass. This correspondence is observed, but not limited to, in clusters HC-A_e and HC-A_c (15.65% and 15.70% of the sample), Elixhauser HC-B_e and Charlson HC-C_c (10.73% and 10.71% of the sample), and Elixhauser HC-C_e and Charlson HB-B_c (16.75% and 16.82% of the sample). Beyond representing similar proportions of the population, these pairs of clusters present virtually the same prevalence of sepsis, mortality and mortality given sepsis (see Figure 4.19). This trend persists at the lower hierarchical level, where each of the clusters in the previously mentioned pairs divides into clusters with comparable size, demographic profiles, and prevalence of outcomes. This similarity in the clustering structure despite differences in morbidity resonates with the findings in Sections 4.3.1 and 4.3.2 underpinning the bias towards the denser features, i.e. demographics and admission type, in detriment of the sparse ones, i.e. comorbidities.

In contrast, the hSBM presents a substantially different picture. As shown in Figures 4.11 and 4.13, the cluster structures inferred for both indices diverge significantly. At an intermediate level, hSBM identifies six clusters for the Elixhauser and only three

for the Charlson index. While some similarities in the clustering structures can be spotted between indices at this level, they fade away as we look into the lower hierarchical level. For instance, hSBM identify clusters of healthy patients under 65 years old in the clusters A_e and C_e (adding up to 19.94% of the sample) and cluster A_c (23.90% of the sample). However, this apparent similarity dissipates when we explore the lower hierarchical level. Here, we can observe that the Elixhauser clusters do not divide further, whereas the Charlson clusters branch into three distinct ones: $A1_c$, comprising patients with the highest prevalence of AIDS and mild liver disease (1,91% of the sample); $A2_c$, encompassing elective patients without comorbidities (3.93% of the sample); and finally $A3_c$, involving non-elective younger patients without comorbidities (18.06% of the sample). These results underscore the hSBM's ability to integrate all features, taking into account demographics when relevant (i.e. younger patients at the intermediate level) as well as morbidities (i.e. clusters $A1_c$, $A2_c$ and $A3_c$).

4.3.3.2 Common clusters found across multimorbidity indices

Surprisingly, hSBM not only identifies similarities based on demographics like the benchmark but also incorporates morbidity count as a distinguishing factor. The robustness of these clusters is further validated by the congruence observed in terms of mortality rates, sepsis onset, and mortality given sepsis among these analogous clusters. In the following, we provide three examples, leaving the remaining comparisons in Annex B.4.

Patients without morbidities (Elixhauser cluster $A1_e$, Charlson cluster $A2_c$ & $A3_c$)

Both hSBM models agree on clusters formed by patients without comorbidities. The first example is that of young patients without morbidities, found in cluster $A1_e$ and cluster $A3_c$. These clusters exhibit a high prevalence of younger patients and low prevalence of elective admission. As expected, they also show significantly lower rates of sepsis (9.68% and 20.07%) and mortality given sepsis (13.3% and 13.63%) compared to the overall averages (27.52% and 21.6%). Another example is cluster $A2_c$, representing **middle-aged patients** without comorbidities with a remarkably low incidence of sepsis onset (5.36%) and mortality given sepsis (3.70%). Notably, cluster $A2_c$ represents 18.06% of the sample being the second biggest cluster in the lower hierarchical level. A common characteristic among all these clusters is their low prevalence of adverse outcomes. The lack of comorbidities distinguishes these clusters aside from

any other cluster in the benchmarks.

Higher morbidity count (Elixhauser cluster F2$_e$ & F4$_e$, Charlson cluster C1$_c$ & C2$_c$)

Cluster F2$_e$ and cluster C1$_c$ show the highest multimorbidity count in their respective morbidity indices, strictly more than 6 and 4 respectively. The high number of comorbidities is reflected in their sepsis prevalence, the highest among all clusters with older patients (56.11% for F2 and 42.22% for C1$_c$), and mortality rate (21.29% and 18.00% respectively). **Cluster F4$_e$ and cluster C2$_c$** present the second-highest comorbidity count among clusters with older patients, 5 or 6 for F4$_e$ and strictly 3 for C2$_c$. These clusters also exhibit high rates of sepsis (roughly 40% for F4$_e$ and 36.23% for C2$_c$) and mortality (18% and 15.80%, respectively), although lower than the previous clusters, consistent with their lower comorbidity count. The high prevalence of negative outcomes is to be expected given the age of patients, their high comorbidity count, and their complex multimorbidity profiles which include a number of cardiovascular diseases (Arnold et al., 2005).

Patients under 85 years old with exactly one multimorbidity (Elixhauser clusters C1$_e$ & Charlson clusters C4$_c$)

These sizeable clusters are characterized for including patients with exactly one comorbidity and excluding patients over 85 years old. Notably, the only difference in outcome is the onset of sepsis 14.27% for C1$_e$ versus 24.27% for C4$_c$ which could be explained by a lower rate of elective admissions in C4$_c$. Interestingly, this does not translate into mortality as its prevalence is lower than the population at 8.25% for C1$_e$ and 9.18% for C4$_c$. This reinforces that the most relevant factors in this cluster are the low comorbidity count and the exclusion of patients over 85.

4.3.3.3 Distinct clusters found between different indices

Of note, hSBM is able to identify clusters that are unique to each morbidity index, underscoring its capability to unveil patient profiles rooted in specific morbidities. The robustness of these multimorbidity profiles is supported by their high correlation with distinct morbidity counts and their associations with mortality, sepsis and mortality-given sepsis rates. In the following, we provide two examples, leaving the remaining

comparisons in Annex B.5.

Patients with mental and neurological disorders (Elixhauser clusters $D1_e$, $D2_e$, $D3_e$)

Cluster D_e is characterized by a higher prevalence of substance abuse among patients aged 25 to 64, along with several related conditions such as AIDS, coagulopathy, liver disease, and fluid-electrolyte disorders (Ballard, 1997; Salmon-Ceron et al., 2005). At the lower hierarchical level cluster D_e splits into clusters $D1_e$ and $D2_e$, showing the highest prevalence of substance abuse, AIDS, and liver disease among all clusters (see Figure 4.11) and cluster $D3_e$ with the highest prevalence of obesity in the sample, 28.22% vs an average of 4.92%. This suggests that a major characteristic of cluster D_e is related to mental health issues (psychosis and depression) connected to drug abuse or obesity. Interestingly, while similar profiles can be seen in the benchmarks, these do not correlate significantly to adverse outcomes. Elixhauser clusters $L4_e$, $K5_e$ and HC-E_e group only young patients with drug abuse but not with no significant prevalence of sepsis, unlike clusters $D1_e$, $D2_e$ and $D3_e$.

Examining $D1_e$ and $D2_e$, we notice that they differ mostly in the multimorbidity count (3 or 4 for patients in $D2_e$ and 5 or higher for patients in $D1_e$). An interesting perspective of these clusters appears when comparing them to Elixhauser clusters $B1_e$ and $B2_e$, which group young patients with substance abuse. Clusters $D1_e$ and $D2_e$ seem to represent and end path for younger patients with substance abuse and low co-morbidity count to older patients with substance abuse complex multimorbidiy profiles (e.g. higher prevalence of liver disease, coagulopathy, peptic ulcer, and weight loss). This is reflected in their higher sepsis and mortality rates (Figure 4.17).

Middle-age patients with exactly one comorbidity, liver disease or aids (Charlson cluster $A1_c$)

This is a very specialized cluster (1.91% of the sample), focused on younger patients with a significantly higher prevalence of severe liver disease (8.6 times above the sample average) and AIDS (11 times above the sample average). Despite their low co-morbidity count (exactly one) and younger age, these patients experience high rates of mortality, sepsis, and mortality given sepsis. Unlike clusters $A2_c$ and $A3_c$, which group similarly young patients with low mortality rates, this cluster identifies patients at high risk of adverse events like sepsis and mortality. Charlson cluster HC-$B1_c$ presents a

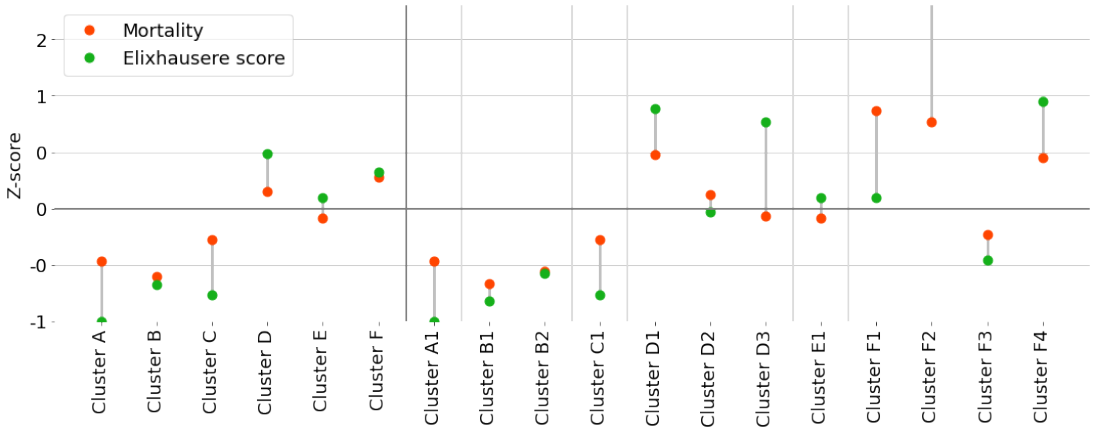

Figure 4.20: z-score for mortality prevalence and Elixhauser score for hSBM clusters.

similar comorbidity profile but doesn't correlate with negative outcome rates as $A1_c$ does. This difference may be attributed to HC-$B1_c$ overemphasizing age, restricting patients to those aged 45 to 65, and underestimating the significance of morbidity count.

4.3.4 Correlation of hSBM clusters and multimorbidity scores

Comorbidity indices offer a scoring system (discussed in Section 2.1.2) to quantify the burden of comorbidity on a patient by assigning a numerical score indicating the impact of each comorbidity on the mortality risk. These scores are particularly valuable as they directly account for multimorbidity and provide insights into the relevance of the multimorbidity profiles we've uncovered.

In this section, we study the relation between the Elixhauser and Charlson morbidity scores and the actual occurrence of mortality within the hSBM clusters. To perform this evaluation, we use the z-scores, a summary statistic that indicates how far from the population mean is a particular observation. In our case we compute the z-score of mortality prevalence with the average Charlson and Elixhauser scores within each cluster. We aim to confirm that the z-score of mortality and indices should be similar to indicate a high correlation between them.

It can be seen that in general, there is a high correlation between the mortality prevalence in the hSBM clusters and the different scores, as presented in Figures 4.20 and 4.21. This is more evident at the higher hierarchical level for both indices, where a significant positive or negative z-score is generally followed by a significantly higher or lower Elixhauser or Charlson z-score. The only exceptions to this at the intermediate

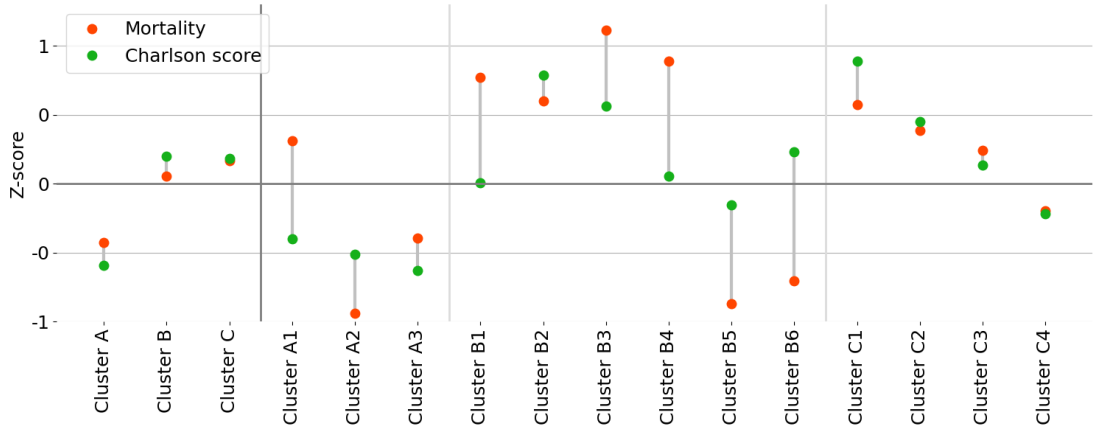

Figure 4.21: z-score for mortality prevalence and Charlson score for hSBM clusters.

hierarchical level are the Elixhauser cluster A_e and C_e (see Figure 4.20). Interestingly, in these cases the corresponding morbidity score underestimates mortality risk. Upon closer examination, we find that these clusters include patients with a low number of comorbidities (strictly 0 for A_e and strictly 1 for C_e). This observation may explain the underestimation, as comorbidity scores are influenced by the number of morbidities. The differences between mortality prevalence and scores are further increased when we delve into the lower hierarchical level of the hSBM-found clusters.

A notable example of these differences is the Charlson cluster $A1_c$, which shows a prevalence of mortality 31% higher than the average population but a Charlson score 40% lower than the average. In this case, we see that the Charlson score underestimates the mortality risk as patients present only one comorbidity and that comorbidity is predominantly mild liver disease, which is considered of little relevance in terms of mortality (weighted of one in the Charlson score). Another example is cluster $B6_c$, presenting a mortality prevalence 70% lower than the average sample, but a Charlson score 23% higher than the average. In this opportunity, the higher Charlson score can be understood due to the relatively high number of comorbidities in patients in this cluster. A final example is clusters $B1_c$ and $B4_c$, both showing a very high mortality prevalence compared to the average population (77.3% and 88.6%) but average Charlson scores. In both cases, we see that the average score can be explained by the moderate number of comorbidities (2 for B1 and 1 for B4). Unfortunately, in these cases, the Charlson score fails to capture the complexity of the profiles of patients in $B1_c$ (middle-aged patients with aids and some life complications) and in $B4_c$ (patients over 85 years old mostly with dementia and with a non-planned admission).

A similar trend can be found in Elixhauser clusters $F1_e$, $F2_e$ and $F4_e$, showing a

mortality prevalence higher than in the average population (87% and 77% and 44.8% respectively), but an Elixhauser score dramatically different between them, 9.3% and 186% and 94.9% higher than the average population respectively. Upon closer examination, is clear that the score is a product of the extremely high number of comorbidities in $F2_e$ (≥ 7) versus $F4_e$ (5 and 6) and $F1_e$ (between 2 and 4). It is noticeable that the Elixhauser score level correlates with the number of morbidities, ranking in descendant order $F2_e$, $F4_e$ and $F1_e$ when the mortality prevalence follows the exact opposite direction. Cluster $D3_e$ is another example of the extreme importance of morbidity count in the Elixhauser multimorbidity score. Despite showing an average mortality prevalence, cluster $D3_e$ shows an Elixhauser score 77% higher than the average population, explained by the high morbidity count in this cluster (≥ 4).

Our analysis uncovers an important observation: while the multimorbidity scores generally correlate with the mortality prevalence, they can overestimate the relevance of morbidity counts in the mortality risk. This underscores the value of hSBM, which is able to simultaneously account for comorbidity count while weighting in relevant relationships between patient's morbidities and demographic characteristics.

4.4 Discussion

Our results highlight our approach's ability to identify complex yet relevant ICU multimorbidity profiles. These profiles interweave demographics and the effects of morbidities, even non-explicit aspects of it such as comorbidity counts. As a result, our approach retrieves known profiles while unveiling new ones, setting our approach apart form benchmark methods. Next, we discuss five key advantages of our approach.

Balance between sparse and dense data: Our results show a balance between profiles defined mostly by demographics and admission type (dense features in our cohort) and morbidities (sparse in our cohort). This balance becomes evident when comparing the clusters identified using benchmark methods across different multimorbidity indices. As pointed out in section 4.2.2, most of these clusters rely on admission type or demographics, the only features consistently present across the various datasets. In contrast, our approach reveals markedly different profiles across datasets utilizing different comorbidity indices.

Uncovering of complex relationships: Our approach is able to incorporate multimorbidity count into the profile inference process, even when it's not an explicit attribute of the dataset. This is in contrast with the benchmark methods that are blind

to this aspect of the data. One example of this is Elixhauser cluster A1 and Charlson cluster A2 and A3, which include patients with no long-term conditions, while this is not captured by any benchmark model. This capacity to capture such unexpected patient heterogeneity is a notable strength of our approach.

Simplified model selection process: Unlike benchmark methods, our use of hSBM allows model parameters to be directly inferred from the data. This is achieved without the risk of overfitting and without the need to account for unbalanced data. This has a potential impact on the usability of our models, as model selection does not play a role in the application of our approach.

Membership to the hSBM profiles as mortality risk indicator: Interestingly, our profiles group patients with similar mortality risks more effectively than the Charlson and Elixhauser scores in our study cohort. This observation can be partially understood by the sophisticated nature of our profiles. In contrast to the morbidity scores, computed as a function of the number of morbidities, our profiles capture the effects of morbidities, demographics and morbidity count. These results evidence the relevance of balancing multimorbidities and demographics for risk assessment in the ICU and highlight the potential of hSBM to complement existing risk assessment methods.

Limitations and future work

While our results are promising, they also open the door to further research aimed at enhancing the validity of our findings and the quality of our models. One limitation of our current work is its reliance on data from a single medical centre. This means that our dataset is affected by patient catchment, hospital protocols, resources and staff characteristics. These factors can impact the generalizability of our results and make external validation challenging. A mitigation for this is the validation of various of the profiles we identified in existing medical literature. Additionally, expanding our research to include other databases will help validate our results and the methodological advantages of our approach more comprehensively.

Another limitation relates to the use of ICD-9 codes, originally designed for billing and administrative purposes rather than detailed clinical descriptions. Furthermore, it is administrative staff rather than medical staff that assign these codes (Johnson et al., 2016a), potentially leading to records emphasising administrative relevant events more than medical ones. This can lead to missing data due to differences in medical versus administrative criteria or incentives to omit certain conditions for administrative or

insurance purposes. Other issues are related to the variability in coding practices in different hospitals. To mitigate these issues, we rely in part on curation of the MIMIC-III as a research tool, which has made it a widely used in the healthcare domain, and the use of non-ICD-9 features such as age and gender. Nevertheless, incorporating additional patient information, such as laboratory results and bedside readings available in the same dataset, can reduce our reliance on ICD-9 codes and offer a more comprehensive view of patient conditions.

The use of the multimorbidity indices also has its limitations, as different indices reflect different medical goals and realities. To account for this we select two widely used indices to demonstrate the robustness of our methodology. However, in future research, we aim to dispense with predefined comorbidity definitions. Instead, we plan to leverage the scalability of network-based algorithms to directly explore disease groupings from the data. This approach can complement medical knowledge with purely data-driven multimorbidity profiles.

An interesting direction for the extension of our work is the incorporation of medical knowledge. Recent literature presents examples of this in the context of community detection (Martin et al., 2016), where medical knowledge is integrated in the form of informative priors for the network parameters. Extending our work in this direction could bridge the gap between medical and data-driven knowledge.

4.5 Conclusion

As presented at the beginning of this chapter, the task of multimorbidity profile identification is of utmost importance in critical care settings. This relevance plus the wealth of ICU data available, like the MIMIC-III dataset, has fueled an active area of work to identify multimorbidity profile patterns in critical care. Not surprisingly, a vast amount of work employs machine learning approaches to this with positive results. Nevertheless, various challenges exist for the use of machine learning in the complex ICU setting. Utilizing our approach we have been able to identify informative profiles while overcoming various limitations that existing methods exhibit.

Besides the contributions outlined in the previous section in the specific multimorbidity profile identification, our results contribute to answering the more fundamental research questions we present in Chapter 1:

- **RQ1**: The combination of all features in the inference of multimorbidity profiles

highlights the ability of our approaches to accommodate and effectively weigh sparser data, unlike the benchmarks. In this way, our approach can identify clusters defined by patient characteristics as well as others defined by multimorbidities. This points to the better adequacy of our approach to ICU data than currently used methods.

- **RQ2**: Notably, our approach was able to unveil and incorporate the effect of comorbidity count into the profiles. This aspect of heterogeneity amongst patients, completely overlooked by benchmark methods, proved itself important especially when looking at mortality prevalence.

- **RQ3**: The non-parametric aspect of our approach effectively removes the need for manual model training and validation. In this way, our approach uses the evidence in the data to simplify the application of machine learning in the ICU.

In conclusion, the research described in this chapter presents a step forward in the use of network sciences for the understanding of multimorbidity profiles in the ICU. Additionally, our approach is able to addresses the challenges the the ICU setting poses to the applications of machine learning presented in Chapter 1. In the next chapter, we lean on our findings to enhance machine learning in the assessment of mortality risk in the ICU.

Chapter 5

Improving mortality prediction for ICU data and patient heterogeneity

The critical conditions of patients combined with the scarcity of resources pose significant challenges to the provision of care in the ICU, as mentioned in Section 1.3.2. In this context, the evaluation of patients based on clinical variables is fundamental to optimising medical interventions (Johnson et al., 2016b). In particular, survival prediction is a critical factor for individual patient assessment and prioritization, and overall ICU performance evaluation (Keuning et al., 2020; El-Manzalawy et al., 2021).

Severity of illness scores assesses patients by quantifying the deviation of clinical variables from normal medical ranges. However, as patient populations and treatments evolve, these scores lose predictive accuracy and calibration. Key limitations include low adaptability to new medical practices (e.g. new treatments), inaccurate model assumptions (e.g. prevalence of conditions), and patient heterogeneity (Le Gall et al., 1993; Zimmerman et al., 2006). Machine learning models for mortality prediction have emerged as alternatives to these severity scores. Yet, as discussed in Section 5.1.2, these models also have drawbacks, such as dependence on data quality, use of outdated patient information, and increased complexity.

Building on our previous findings, we explore the use of network modelling and hierarchical Stochastic Block Modeling to tackle these limitations. We use hierarchical community detection to extract relevant relationships between groups of clinical features and patients. Leveraging this block structure, we use machine learning to build mortality prediction models at each hierarchical level. Our approach provides a robust and flexible scoring system that is adaptable to varying data quality and shows consistent top predictive performance across different age and ethnicity cohorts.

5.1 Mortality prediction in the ICU

As discussed in 1.3.2, the accurate prognosis of ICU patient mortality is a critical task in the ICU. Traditionally, severity scores serve this purpose, focusing on key clinical variables to represent a patient's health status and associate its deviation from normal ranges with the probability of mortality. However, scoring systems show declining performance over time in the ICU. To address this, machine learning has been explored in recent research. These models show improved performance but remain sensitive to missing data and introduce complexity. In this section, we discuss existing approaches for mortality prediction in the ICU and introduce our own.

5.1.1 Scoring systems

Over the past decades, various scoring systems have been developed to assess the severity of a patient's condition based on clinical variables collected within the initial 24-48 hours after admission (Keuning et al., 2020). Two of the advantages of these methods are their simplicity of use and the provision of an intuitive understanding of a patient's condition (El-Manzalawy et al., 2021).

These models, discussed in Section 2.1.4, involve the discretization of clinical variables (e.g. heart rate or temperature) into different ranges. Each range is assigned a numerical value, called a subscore, indicating its deviation from clinically normal ranges. For instance, a heart rate subscore of *0* signifies readings within the normal clinical range, while higher subscore values indicate increasing degrees of deviation. These are then summed up to calculate a score that is associated with the patient's overall mortality risk (Vincent and Moreno, 2010).

These systems combine medical knowledge, statistical techniques and machine learning models in various ways to select the clinical variables of interest and define the relevant ranges and subscores to capture their impact on the patient's condition. Despite their methodological variation, all severity scores finally present a simple scoring card used to evaluate patients. In Table 5.1 we show the variables for SAPS-II, which will be used as part of our work in Section 5.2.4.

To illustrate, let us consider a 43-year-old patient admitted to the ICU for an unscheduled surgery. In Figure 5.1, we summarize the patients' worst values for the relevant SAPS-II clinical variables within the first 24 hours from admission. In SAPS-II, *worst* refers to the most extreme value for a variable. For variables such as body temperature, it means the highest temperature registered, for others like the Glasgow

Variable	Range	Subscore	Variable	Range	Subscore
Age,	<40	0	WBC Count (WC),	1.0-19.9	0
years	40-59	7	10_3/cu mm	>20.0	3
	60-69	12		<1.0	12
	70-74	15	Serum Pottasium (SP),	3.0-4.9	0
	75-79	16	mmol/d	<3.0	3
	>80	18		>5.0	3
Heart rate (HR),	70-119	0	Serum Sodium level (SS),	125-144	0
beats/minute	40-69	2	mmol/L	>145	1
	120-159	4		<125	5
	>160	7	Serum bicarbonate level	>20	0
	<40	11	(SB), mEq/L	15-19	3
Systolic Blood Preasure (BP),	100-199	0		<15	6
mm Hg	>200	2	Bilirubin level (BL),	<68.4	0
	70-99	5	micro mol/L	68-102.5	4
	<70	13		>102.6	9
Body Temperature (T),	<39	0	Glasgow Coma Score	14-15	0
°C	>39	3	(GCS)	11-13	5
Pulmonary Artery Preassure	100-199	9		9-10	7
(PA), mm Hg/FIO2	>200	6		6-8	13
if Ventilated	<100	11		<6	26
Urinary output (UO),	>1.000	0	Chronic diseases	Metastatic cancer	9
(lt/day)	0.500-0.999	4		Hematologic malignancy	10
	<0.500	11		AIDS	17
Serum Urea level (SU),	<10	0	Type of admission (TA)	Schedule surgical	0
(g/L)	10.0-29.9	6		Medical	6
	>30	10		Unschedule surgical	8

Table 5.1: SAPS-II scores mapping to clinical variable as in Le Gall et al. (1993)

Coma Score (GCS) it means the lowest recorded value, and for others like heart rate, the most extreme (either higher or lower). Using the data, we compute the SAPS-II score by adding up the subscores (from the scoring card in Table 5.1), resulting in a score of 71 points for this patient. With this score, we calculate the mortality probability using the SAPS-II transformation function, obtaining a mortality probability of 0.85. Note that in this example, the subscore related to Serum Potassium is considered as *0*, assuming the missing potassium value is within normal ranges.

	Age	Hear rate	Syst.Blood Preasure	Body Temperature	Pulmonary Preasure	Urinary output	Serum Urea Level	WBC	Serum Pottasium
Value	43	160	90	42	210	0.4	9.8	0.8	Not recorded
Subscore	7	7	5	3	6	11	0	12	0

	Serum Sodium	Serum bicarbonate	Bilirubin level	Glasgow Coma Score	Chronic disease	Type of Admission
Value	150	14	66	12	Not recorded	Unschedule surgical
Subscore	1	6	0	5	0	8

SAPS-II score: 71 points

Mortality probability: 85.0%

Figure 5.1: Computation of SAPS-II scores and mortality probability. Missing values are assumed in normal clinical ranges, and its associated subscore is imputed as **0**.

While simple in use, the performance of these severity scores declines over time, as

noted in studies reporting reduced discriminative power (Zimmerman et al., 2006), underestimated mortality rates (Metnitz et al., 2005), and loss of output calibration (Walsh et al., 2017). This decline also affects patient subgroups, e.g. defined by age or ethnicity (Moreno and Apolone, 1997; Strand et al., 2009). Limited adaptability to changing medical conditions has been cited as a reason for this decline (Metnitz et al., 2005). For example, new medical protocols and treatments introduced after the severity score is established can significantly improve the prognosis of patients, rendering it outdated. Patient heterogeneity is another problem for severity scores (Strand et al., 2009; Zimmerman et al., 2006). Important patterns of clinical variables specific to underrepresented groups of patients can be overlooked as predictive models focus on the larger patient groups. Finally, inaccurate model assumptions, such as variable independence between predictive variables, contribute to the issue (Le Gall et al., 1993; El-Manzalawy et al., 2021).

5.1.2 Machine learning models for patient mortality prediction

Machine learning algorithms like Random Forest (Breiman, 2001) (RF), eXtreme Gradient Boosting (XGB) (Chen and Guestrin, 2016), and Artificial Neural Networks (Rumelhart et al., 1985) have been proposed in response to the limitations shown by traditional severity scores (Kong et al., 2020; El-Rashidy et al., 2020). These algorithms aim to address these limitations by capturing non-linear relationships between mortality and clinical variables. While they show promising results, they are often tailored to specific patient cohorts (Lin et al., 2019; Carioli et al., 2020), which can limit their generalizability. Furthermore, they often use data that differs from that of established severity scores, which makes its application in clinical settings difficult. Additionally, the inherent lack of transparency in machine learning models can affects their practical application (El-Manzalawy et al., 2021).

Another approach is to leverage machine learning techniques to develop predictive models that build upon existing scoring systems (El-Manzalawy et al., 2021; Pearce et al., 2006). A prominent example is OASIS+ (El-Manzalawy et al., 2021), which uses OASIS subscores (see Section 2.1.4) to propose a severity of illness aligned with a validated system while benefiting from machine learning. Although OASIS+ shows improved prognostic performance and output calibration, it requires the same information as OASIS and depends on the quality of OASIS scores and imputation methodology. We further discuss OASIS+ in the context of our work in Section 5.2.3.

5.1.3 Hierarchical stochastic block modelling enhanced models

In this section, we present an alternative approach to leverage the existing severity of scores and machine learning models building on the benefits of community detection in networks. While we are aware of the use of networks in the context of multimorbidity analysis, these are used for descriptive analysis of networks and clustering, such as in the effects of sex (Kalgotra et al., 2017) and ethnicity (Kalgotra et al., 2020) in multimorbidity patterns. This is how to the best of our knowledge our work represents the first use of network science for severity risk assessment in the ICU. Next, we present an overview of our methodology, leaving the details for Section 5.2.4.

We start by using the defined clinical variable and ranges of SAPS-II to build a bipartite network of patients and their clinical information. By doing this, we align our model with SAPS-II, a clinically validated and widely used system that works with routinely captured information in the ICU (Le Gall et al., 1993; Kong et al., 2020). Notably, we do not use SAPS-II subscores in our network, dismissing their role and relying on our models' ability to capture relationships from the graph. Then, we use hSBM to infer a hierarchical block structure, see Figure 5.3. We finally use machine learning to create mortality predictive models using the connections between a patient and the blocks as input, as presented in Section 5.2.4. By doing this we effectively perform a dimensional reduction from the SAPS-II subscore space into a SAPS-II block space. In Section 5.3.3 we show that this step is crucial to our model's ability to handle missing information automatically without making assumptions about the reasons for the missing data.

We aim to exploit the resilience of our approach to reexamine the need for imputation that current models present. Moreover, we expect that the ability of our model to capture and model patient heterogeneity will help to tackle the lack of performance across different patient cohorts. Finally, we expect to lean on our models' ability to automatically adapt to the data to achieve all of these expected benefits without heavy human interventions.

5.2 Experimental setting

Figure 5.2 outlines our experimental pipeline, illustrating the three main sets of experiments conducted to assess our approach's effectiveness for mortality risk assessment. We start by exploring the performance of widely used severity of scores (1). Then we

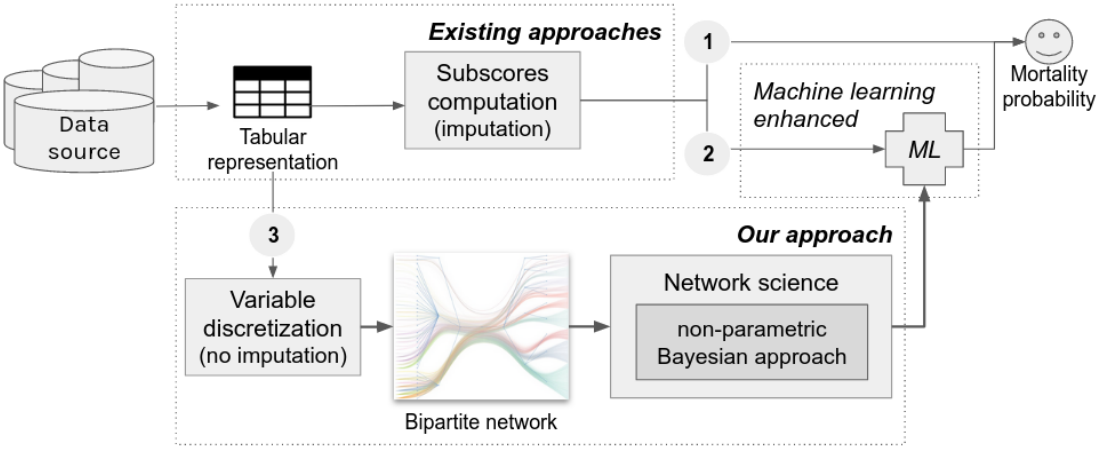

Figure 5.2: Experimental pipeline overview. The severity score (1) and machine learning enhanced (2) experiments, depicted on the top of the diagram, constitute our benchmark. Our approach (3), involves network-based data modelling and the automatic detection of feature communities prior to the use of machine learning.

look into machine learning-enhanced models (2). These models are OASIS+ and XG-SAPS, treated in more detail in Section 5.3.2. Finally, we utilize hXG-SAPS (3), our proposed model described in Section 5.2.4. All experiments make use of the patient cohort described in Section 3.4.

5.2.1 Patient cohort for mortality prediction study.

As a reminder of our study cohort, detailed in Section 3.4, includes patients aged 18 to 90 with ICU stays lasting at least 24 hours, considering only the first admissions for those with multiple ones. It comprises 31,908 unique patients, their corresponding ICU stays, clinical variables, and scores associated with the five severity scores in this study (see Table Table 3.1). On average, patients are 62.77 years old, with an average length of stay of 4.7 days. Males make up 57,89% of the sample. The cohort shows an imbalanced mortality rate, with 10.57% of patients dying in the ICU.

To perform the analysis, we considered patient age into clinically relevant categories (Barnett et al., 2012c; Zador et al., 2019): 18-24, 25-44, 45-64, 65-84, and over 85 years old ranges. Patient ethnicity follows the MIMIC-III dataset categories (Johnson et al., 2016a): White, Black, Asian, Hispanic, and Other. Additionally, we incorporated the probability of mortality as defined by each model when available (SAPS-II, APS-III, and OASIS). In cases where these probabilities were not provided (SOFA and SIRS), we calculated a pseudo-probability by normalizing patients' scores

Models		Model input	Data processing	Output computation
(1) Severity of illness scores	OASIS, SAPS-II, etc…	Clinical variables	Subscores	Subscores summation
(2) Machine learning models	OASIS+	OASIS Subscores	------	Predictive model
	XG-SAPS	SAPS Subscores	------	Predictive model
(3) hSBM enhanced models	weighted hXG-SAPS	SAPS Subscores	Weighted feature blocks	Predictive model
	unweighted hXG-SAPS	SAPS Subscores	Unweighted feature blocks	Predictive model

Table 5.2: Mortality prediction models. Input, a short description of any data processing done before the prediction, and the method used to perform the mortality prediction are presented for each model. "Subscores" indicate that clinical variables are discretized into subscores. The number next to the model names associates the models with the pipeline paths presented in Figure 5.2.

within the range of [0,1], as discussed in Section 3.4.

As discussed in Section 3.3, subscores corresponding to missing clinical variables are imputed as *0* in all severity scores, assuming readings within normal range. However, for our network-based model, we do not impute scores (see Section 5.2.4). Specifically, for SAPS-II subscores, 58.33% of patients lacked pulmonary artery pressure data, and 55.96% lacked bilirubin level information. The remaining SAPS-II variables (in Table 5.3) had lower rates of missing data: 1.04% for heart rate, 1.12% for systolic blood pressure, 2.92% for temperature, 3.12% for urine output, 0.82% for serum urea level, 1.28% for white blood count (WBC), 0.71% for serum potassium, 0.81% for serum sodium, 1.29% for serum bicarbonate, and 1.08% for GCS.

For supervised learning, we randomly divided the study cohort into training and test datasets with a 70% - 30% split. We ensured representative datasets by maintaining an equal distribution of features as in the entire study cohort (chi-square tests with p-values ≤ 0.05). Table 3.1 shows that both datasets have nearly identical average patient age (62.8 years old), length of stay (4.7 days), mortality prevalence (approximately 10,5%), and gender composition (close to 42% females across datasets). The machine learning enhanced models in Section 5.2.3 and our models in Section 5.3.3 were trained and tested using these train and test datasets.

5.2.2 Severity of illness scores for mortality prediction

Our initial experiments focused on using severity scores as predictors for mortality (see Figure 5.2, 1). Specifically, SAPS-II, OASIS, APS-III, SOFA and SIRS, which are al-

ready available in MIMIC-III and presented in Section 3.4. However, for evaluating mortality prediction performance, we need to utilize the models' mortality prediction instead of the severity scores. These probabilities are calculated using a transformation provided as part of the scoring system in SAPS-II (Le Gall et al., 1993), APS-III (Zimmerman et al., 2006), and OASIS (Johnson et al., 2013a), readily available in MIMIC-III. For SOFA and SIRS, lacking such probabilities, we derived pseudo-probabilities by normalizing patients' scores to the [0,1] range, as discussed in Section 3.4.

Our final prediction models for each severity score take a patient's clinical variables during the first 24 hours of admission as input and convert them into a severity score, as exemplified in Section 5.1.1. Finally, these scores are transformed into mortality probabilities, as described earlier. Following this methodology, we created five mortality prediction models using the SAPS-II, OASIS, APS-III, SOFA and SIRS. Table 5.3 shows the performance of these methods, discussed in Section 5.3.1.

5.2.3 Machine learning enhanced models for mortality prediction

In our second set of experiments, we explored the use of machine learning to enhance OASIS and SAPS-II, two of the best-performing severity scores as discussed in Section 5.2.3. Following existing literature, we replicated OASIS+, employing OASIS subscores as features to train an XGBoost model with 200 trees. As shown in Figure 5.4, our final model demonstrated performance consistent with literature reports (AUROC: 0.80) (El-Manzalawy et al., 2021).

As see in Figure 5.4, despite having and improved performance, OASIS+ only achieves the performance of SAPS-II (AUROC: 0.80). The outstanding results of SAPS-II inspired us to expand on this approach and to take on the challenge of further improving its performance. Similar to OASIS+, we utilized SAPS-II subscores (see Table 5.1) for each patient, training an XGBoost model (XG-SAPS) and a Random Forest one (RF-SAPS). Both models were ensembles of 200 trees to maintain consistency with OASIS+. Upon comparing their performances, we opted to exclude RF-SAPS as a benchmark due to its consistent underperformance in mortality prediction. The results are provided in Figure C.4 of ther Appendix C .

Our final machine learning enhanced models are the replicated **OASIS+** and **XG-SAPS**. Each receives patient's subscores (OASIS and SAPS) as input and generates a mortality prediction using an XGBoost model. It is important to note that, as discussed in Section 3.3, the subscores serving as input for these models are imputed with a value

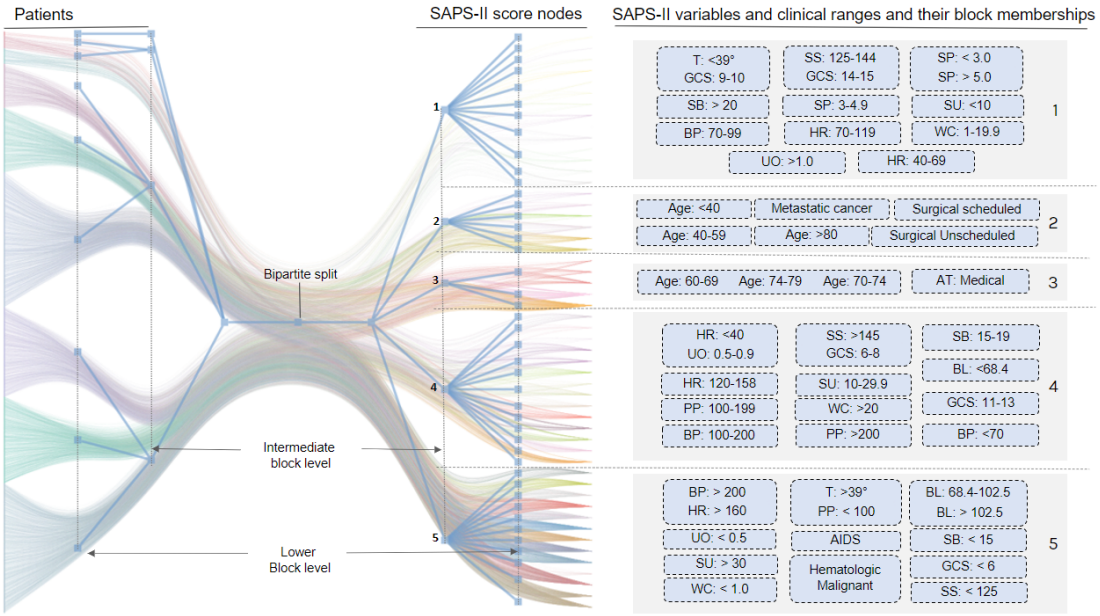

Figure 5.3: The network between patients and features is illustrated by the bipartite network on the left. The inferred hierarchical block structure is depicted by the tree on top of the network. Each feature block and its block membership are shown on the right. Light blue boxes group nodes with the same lower block membership and grey ones depict their intermediate block membership. For example, the “60-69”, “70-74” and “75-79” age nodes are grouped together in the same block at the lower level, and together with the ”Medical” Admission Type in block 3 at the intermediate block level.

of zero if there is no clinical record to calculate it. Table 5.2 summarizes the inputs and outputs of these models.

5.2.4 Network enhanced models for mortality prediction

In this section, we present our mortality prediction approach, as illustrated in Figure 5.2 (3). Our methodology involves three steps: inference of the hierarchical clustering model on the patient-clinical variable network, use of the clustering structure to create the patient-cluster matrix, and training of the mortality prediction. We apply this approach to the SAPS-II scoring system due to its widespread use in clinical and research domains and its exceptional performance, as demonstrated in Section 5.3.1.

Inference of the hierarchical clustering model

We begin by constructing a bipartite network, including 53 feature nodes representing each of the SAPS-II subscores (in Table 5.1) and 22,335 nodes, each representing a

patient in our study cohort. Each patient node is connected to as many subscores nodes based on their clinical history. For example, a 43-year-old patient with a temperature of 41°C would link to the score node for their age group (40-59 years old) and to the one for their temperature (over 39°C). We then, employ hSBM to infer a cluster structure for patients and feature nodes using the two-step methodology outlined in Chapter 2.3.3.

The final cluster structure for our cohort is presented in Figure 5.3, depicted by the tree-like structure overlaid on the network. Starting from the ”Bipartite split,” the left branches depict patient clusters and their hierarchy, while the right branches show the hierarchical structure of SAPS-II nodes. The SAPS-II nodes are listed under the SAPS-II variables and clinical ranges column in Figure 5.3, these 52 nodes represent each clinical variable range defined in the score. These nodes are grouped into 43 clusters at the lower block level, shown by the light blue boxes around the initial 52 nodes in Figure 5.3. These 43 clusters are also seen as the leaves of the right branches in the tree-like structure. These clusters are further grouped into 5 intermediate-level clusters.

It is important to note that we do not impute missing values during the network construction. If a patient lacks a record for a variable, then the patient will not be linked to a node representing that missing variable. By doing this, we let the model handle missing data without making assumptions about it. Additionally, unlike SAPS-II, we consider each node as equally important to the patient's condition. We do this by not considering the weights defined in SAPS-II to assess each clinical range, resulting in our unweighted bipartite network to model patients and clinical features. Additionally, while our feature networks represent the SAPS-II clinical variables and their ranges, we treat each node as equally important to the patient's condition. We achieve this by not using the feature subscores as weights for the network links. By doing this, we move our model away from specific calibration parameters used to derive SAPS-II (done on a different cohort than ours), while retaining the medical knowledge behind the definition of the relevant SAPS-II variable.

The patient-cluster membership matrix

In the second step, we leverage the cluster structure to reshape the feature space for training our predictive models. Our aim is to simplify the clinical variables space by making use of the fact that nodes in the same hSBM cluster show the same probability of connecting to other nodes in the network (discussed in Section 2.3.3). In other

words, we can capture relevant patient-variable relationships by focusing on connections between patients and clusters rather than to nodes.

Considering the blocks in Figure 5.3, we create a patient-block matrix to capture the connections of each patient to each block. For instance, at the intermediate block level with five blocks in Figure 5.3, the patient-block matrix has five columns (one for each block) and rows equal to the number of patients in our cohort. In the network, a patient with a body temperature of 37°C will have a link to the block containing the node "Temperature under 39°C" (range defined in SAPS-II). This link is represented in the matrix with a value 1 in the column corresponding to block 1, which includes the "Temperature under 39°C" node.

When considering all the features of a patient, it is possible for a patient to have more than one link to a block. We create two types of patient-block matrix to account for this: the **weighted** matrix, where a column value equals the number of links between a patient and the block, and the **unweighted** matrix, where the block value is set to 1 if there is at least one link between the patient and a node in the block. Additionally, the patient-block matrix can be generated at the intermediate level (as in the previous example) and at the lowest block level, considering 42 blocks in our case. Following the previous approach we create four different matrices: a weighted and an unweighted patient-block matrix at the lower block level, and a weighted and an unweighted patient-block matrix at the intermediate block level.

Training of the mortality prediction

The final step is the training of the mortality predictive model. Maintaining consistency with the previously described machine-learning enhanced models (as detailed in Section 5.2.3), we train four XGBoost models using the patient-block matrices described before. Bellow we describe the final predictive models:

- **Weighted hXG-SAPS:** XGBoost model trained with the weighted patient-block matrix, considering blocks at the lower block level. This model incorporates the most information, considering all connections between patients and the 42 nodes of the lowest block level. Referred to as **hXG-SAPS (w)**.

- **Unweighted hXG-SAPS:** XGBoost model trained with the unweighted patient-block matrix, considering blocks at the lower block level. Referred to as **hXG-SAPS (u)**.

- **Weighted Simplified hXG-SAPS:** XGBoost model trained with the weighted patient-block matrix, considering blocks at the intermediate block level. This model significantly simplifies the input space by considering only 5 blocks instead of 42. Referred to as **(S) hXG-SAPS (w)**.

- **Unweighted Simplified hXG-SAPS:** XGBoost model trained with the unweighted patient-block matrix, considering blocks at the intermediate block level. This is the simplest model, as it considers the reduced 5-block input space and only the existence of a connection. Referred to as **(S) hXG-SAPS (u)**.

5.2.5 Performance metrics for evaluation

We use several metrics to comprehensively assess the predictive performance of severity scores and predictive models. These are the Area under the Receiver-Operating Curve (AUROC), accuracy (ACC), sensitivity (SN), specificity (SP), Observe-to-Expected ratio (O/E), and the root square mean error (RMSE) for the calibration curve of the model's output. A description of these metrics is provided in Chapter 2.5.2.

Out of these metrics, accuracy, sensitivity, and specificity rely on a threshold to transform our models into binary classifiers. Since this threshold is specific to the sample and may be affected by the presence of imbalanced data, we conduct a bootstrapped evaluation to obtain robust estimates of these metrics and enhance the reliability of our analysis. The optimal binarization threshold for each bootstrapped model was computed by minimizing Youden's J statistic, presented in Section 2.5.2. Additionally, we use the AUROC to evaluate the discriminate power of a model. Some of the benefits of AUROC are its robustness to imbalanced data and independence from thresholds (Fawcett, 2006).

A model's output calibration is commonly used to evaluate the quality of the predicted prevalence of mortality at a population level. The Observed-to-Expected ratio (O/E), calculated as the total number of observed deaths divided by the sum of the predicted probabilities of death, is widely used to assess the mortality predicted by a model for a group of patients (Strand et al., 2009; Zimmerman et al., 2006). Additionally, other studies utilize the RMSE of the calibration curve to study the quality of risk stratification of a model (El-Manzalawy et al., 2021; Walsh et al., 2017). In our study, we employ both of these techniques as they provide complementary views on a model.

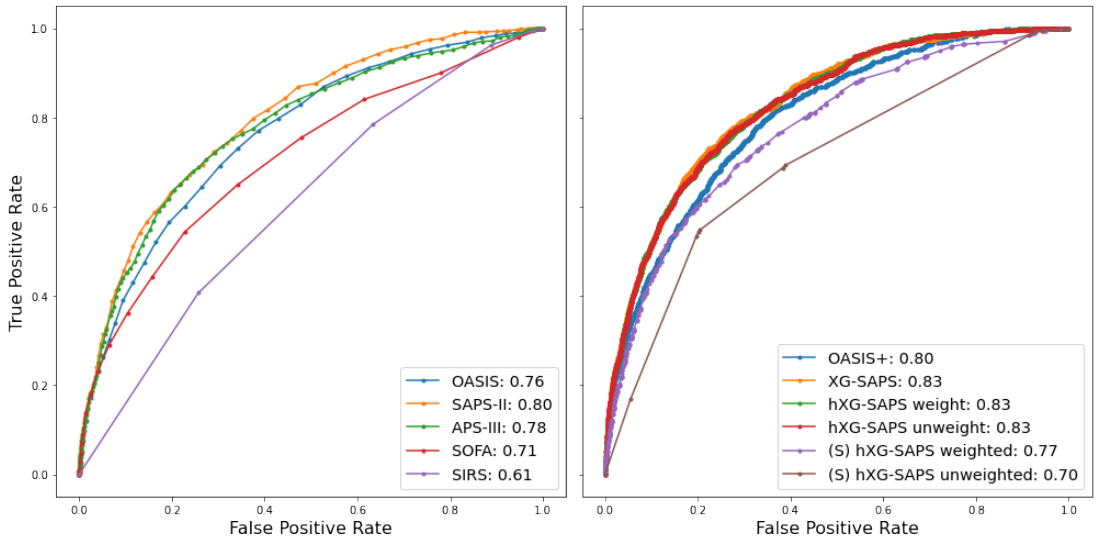

Figure 5.4: AUROC comparison across prediction model. Enhanced methods (right) consistently outperformed severity scores (left). SAPS-II and the hXG-SAPS show the highest performance followed by OASIS+. The simplified (S) hXG-SAPS models show performance comparable to severity scores despite considering a reduced input space.

5.3 Mortality prediction analysis

In this section, we conduct a detailed analysis of our approach and benchmark models, dividing it into three subsections. We first present the performance of the severity scores in Section 5.3.1. Next, in Section 5.3.2, we compared the machine learning enhanced models (OASIS+ and XG-SAPS) with the severity scores and asses the improvements achieved. Subsequently, in Section 5.3.3, we compare the performance of our approach (hXG-SAPS) with the machine learning enhanced modes and severity scores. The presentation of our approach is split into two parts. We begin by assessing hXG-SAPS in Section 5.3.3.1 and its simplified version in Section 5.3.3.2 across the entire study cohort. Finally, we examine the performance of our approach across age and ethnicity cohorts in Sections 5.3.4.1 and 5.3.4.2.

5.3.1 Severity scores performance

We start our discussion by providing an overview of the performance of traditional severity of illness scores for mortality prediction. Our goal is to set a clear baseline for the evaluation of the enhanced models presented in Sections 5.3.2 and 5.3.3.1, describe their performance and validate our study cohort against results in the literature.

Figure 5.4 presents a comparison of the performance of severity scores. It is no-

ticeable how SAPS-II shows the highest discriminant performance (AUROC: 0.80) followed by APS-III, and OASIS (AUROC: 0.78 and 0.76 respectively), and finally SOFA and SIRS (AUROC: 0.71/0.61). Additionally, as shown in Table 5.3 SAPS-II and APS-III keep a high accuracy as well (0.74 and 0.75), balancing sensitivity (0.73 and 0.71) and specificity (0.75 and 0.76). OASIS and SOFA exhibit lower accuracy (ACC: 0.70 and 0.71) potentially indicating overfitting as reflected by their sensitivity and specificity. While OASIS seems to overestimate mortality (SN: 0.73 and SP: 0.69), SOFA seems to underestimate it (SN: 0.61 and SP: 0.73). SIRS consistently ranks as the poorest score in all three metrics (ACC: 0.56, SN: 0.63 and SP: 0.55). When assessing the calibration curves (see Table 5.3), APS-III 0.29, OASIS 0.30 present the best performance (RMSE: 0.29, 0.30), followed by SAPS-II and SOFA (RMSE: 0.32 for both) and finally SIRS (RMSE: 0.70). Notably, APS-III stands out from the other scores with an excellent Observed-to-Expected ratio (O/E: 0.99) compared to OASIS (O/E: 0.74), SAPS-II (O/E: 0.50), and SOFA (O/E: 0.50).

As presented, SAPS-II and APS-III consistently show top performance amongst severity scores. However, the substantial number of required clinical variables to compute them presents a drawback (El-Manzalawy et al., 2021). In the next section, we show how our approach can leverage SAPS-II and address this limitation, achieving state-of-the-art performance with a simpler model.

5.3.2 Machine learning enhanced models performance

In this section, we evaluate the performance of the machine-learning-enhanced models, OASIS+ (El-Manzalawy et al., 2021) and XG-SAPS. These models, presented in Section 5.2.3, are built on top of the subscores of OASIS and SAPS-II, with the goal of improving their performance (El-Manzalawy et al., 2021). In the following, we assess the impact of machine learning on OASIS and SAPS-II in our dataset and establish a benchmark for our network-enhanced model presented in the next section.

A first point to notice from Figure 5.4, is that both machine learning models outperform their base models, with XG-SAPS achieving an AUROC of 0.83 (versus 0.80 for SAPS-II) and OASIS+ reaching 0.80 (versus 0.76 for OASIS). As shown in Table 5.3, these results extend to accuracy, sensitivity, specificity and calibration, with machine learning models surpassing their base severity scores predictive performance.

However, despite improving on their base model, XG-SAPS compares more favourably to the rest of the severity scores than OASIS+. For example, XG-SAPS achieves the

highest AUROC of all models at 0.83, with OASIS+ ranking second, as high as SAPS-II (0.80). The same trend is observed in terms of accuracy, where the best models are XG-SAPS, SAPS-II and APS-III (ACC: 0.76, 0.74 and 0.75). Here, OASIS+ shows an accuracy of 0.71, matching OASIS with 0.70. Moreover, OASIS+ seems to replicate the tendency of overestimate mortality seen in OASIS. OASIS+ shows a high sensitivity (0.77) compared to its specificity (0.70), similar to OASIS (SN: 0.73 and SP: 0.69). In contrast, XG-SAPS shows more balance between its sensitivity (0.75) and specificity (0.77). Furthermore, SAPS-II and APS-III show higher accuracy (ACC: 0.74 and 0.75) and balance between sensitivity (SN: 0.73 and 0.71) and specificity (SP: 0.75 and 0.76) than OASIS+.

A similar trend is observed in terms of output calibration. While the calibration curves of XG-SAPS and OASIS+, along with APS-III, present the best results (RMSE: 0.28, 0.29, 0.29), the Observed-to-Expected reveals shows differences in performance. Here, the best models are XG-SAPS and APS-III (O/E: 1.03 and 0.99), followed by OASIS and OASIS+, which are significantly lower (O/E: 0.74 and 0.73).

These results suggest that machine learning-enhanced models can improve the performance of severity scores. This is shown by the better performance of XG-SAPS over SAPS-II and OASIS+ over OASIS. However, we notice that our new model XG-SAPS significantly outperforms OASIS+ by El-Manzalawy et al. (2021). In the next section, we present our approach to further enhance the performance of XG-SAPS.

5.3.3 hXG-SAPS performance assessment.

We begin the analysis of our approach by examining the clustering structure inferred by hSBM. As shown in Figure 5.3, at the lowest block level the initial 52 nodes are clustered into 43 distinct ones, further aggregating into 5 at the intermediate block level. Notably, 35 of these 43 blocks consist of only one node, indicating a high degree of independence among the ranges of clinical variables defined in SAPS-II. The remaining blocks cluster multiple nodes, suggesting that nodes within these blocks share roles in the network structure and could be considered as one.

For example, at the lower level, two blocks group multiple ranges of the same variable, suggesting a potential to merge these ranges: the 60-69, 70-74, and 75-79 age groups, and bilirubin levels out of the normal range, 68.4-102.5 and $>$ 102.6 μmol/L. Other blocks group different variables together, suggesting redundancy in the subscores. For instance, sodium, Glasgow Coma Scale and heart rate are grouped together

Method	AUROC	Accuracy	Sensitivity	Specificity	RSME	E/O ration
SAPS-II	0.80	0.74	0.73	0.75	0.32	0.50
OASIS	0.76	0.70	**0.73**	0.69	0.30	0.74
APS-III	0.78	**0.75**	0.71	0.76	0.29	0.99
SOFA	0.71	0.71	0.61	0.73	0.32	0.5
SIRS	0.61	0.56	0.63	0.55	0.70	0.15
XG-SAPS	**0.83**	**0.76**	**0.75**	**0.77**	**0.28**	**1.03**
OASIS+	0.80	0.71	0.77	0.70	**0.29**	0.73
hXG-SAPS (w)	**0.83**	**0.75**	0.74	**0.76**	**0.28**	**1.03**
hXG-SAPS (u)	**0.83**	**0.75**	**0.75**	0.75	**0.28**	**1.03**
(S) hXG-SAPS (w)	0.77	0.74	0.67	0.75	**0.29**	**1.04**
(S) hXG-SAPS (u)	0.70	0.77	0.55	0.80	0.30	**1.03**

Table 5.3: Performance of all predictive models applied to the whole study cohort. We report the mean of the bootstrapped evaluation accuracy, sensitivity, and specificity. In bold we highlight the top 3 best performing models for each evaluation metric. Figure C.1 in Appendix C shows the results including the confidence interval.

in blocks such as the lowest sodium level and the highest heart rate, normal levels of sodium and consciousness (15 GCS), and very high levels of sodium and grouped with low consciousness (6-8 GCS). Similar patterns of aggregation are observed in the 5 blocks of the intermediate level with block 3 clustering nodes of age groups between 60-79 years while block 2 encompasses all other age groups, and block 5 group nodes represent normal ranges for all clinical features except for bilirubin.

We capture these patterns using the patient-block matrices presented in Section 5.2.4 as we consider the connections between patients and blocks rather than to other clinical variable nodes. In the following, we show how these matrices allow us to generate models that benefit from these patterns.

5.3.3.1 hXG-SAPS overall performance

As shown in Table 5.3, hXG-SAPS demonstrates performance on par with the best models, despite using less information. Moreover, it ranks as the best model along with XG-SAPS in terms of predictive performance (AUROC: 0.83) and with APS-III in terms of calibration of results (RMSE: 0.29 and O/E: 0.99).

Interestingly, the performance of hXG-SAPS remains consistent between the unweighted and weighted versions (AUROC: 0.83 and ACC: 0.75), underscoring the importance of the link to a feature block rather than the strength of the link. Moreover, the sensitivity (0.74 and 0.75) and specificity (0.76 and 0.75) are virtually the same for

Method	18-24		25-44		45-64		65-84		Over 85	
	AUROC	O/E	AUROC	O/E	AUROC	O/E	AUROC	O/E	AUROC	O/E
SAPS-II	0.86	0.19	**0.86**	0.28	**0.80**	0.30	**0.76**	0.39	0.73	0.48
OASIS	0.84	0.08	0.78	0.14	0.76	0.19	0.74	0.28	**0.74**	0.34
APS-III	0.87	0.17	0.84	0.29	0.78	0.37	0.75	0.57	**0.76**	0.73
SOFA	0.78	0.23	**0.83**	0.33	0.73	0.38	0.67	0.58	0.65	0.80
SIRS	0.73	0.04	0.62	0.08	0.64	0.12	0.61	0.19	0.55	0.26
XG-SAPS	**0.91**	**0.63**	**0.90**	**0.81**	**0.84**	**0.93**	**0.80**	**1.08**	**0.74**	**1.04**
OASIS+	**0.96**	**0.61**	**0.83**	0.65	0.79	0.65	**0.77**	0.77	**0.79**	0.85
hXG-SAPS (w)	**0.86**	**0.73**	**0.89**	**0.87**	**0.84**	**0.91**	**0.80**	**1.17**	0.73	**1.07**
hXG-SAPS (u)	**0.88**	**0.77**	**0.89**	**0.88**	**0.84**	**0.91**	**0.80**	**1.11**	0.73	**1.06**
(S) hXG-SAPS (w)	0.78	0.33	**0.86**	0.35	**0.80**	0.77	**0.76**	1.30	0.73	1.68
(S) hXG-SAPS (u)	0.74	0.33	0.77	0.56	0.74	**0.78**	0.66	1.25	0.63	1.76

Table 5.4: AUROC and Observed-to-Expected ration summary of all predictive models applied to age cohorts. In bold we highlight the top 3 best performing models for each evaluation metric in each cohort.

both versions. Additionally, Table 5.3 further confirms their consistency in terms of calibration, with both showing the same RMSE: (0.28) and O/E ratio (1.03).

These findings suggest that hSBM effectively captures SAPS-II subscores information through block membership, represented in our patient-block matrix (see Section 5.2.4). Furthermore, hXG-SAPS performs consistently across both versions, suggesting that a patient's severity is primarily determined by their association with a block (i.e., having at least one link to such a block) rather than the strength of that association (e.g. the number of links to that block). By doing this, our approach can automatically capture the importance of clinical variables in the block assignment without the need to fit subscores.

5.3.3.2 Simplified hXG-SAPS overall performance

The performance of the simplified hXG-SAPS (see Table 5.3), varies between versions. Notably, the weighted version achieves an AUROC of 0.77 and 0.70 for the unweighted version. However, the unweighted version shows better accuracy (0.77) than the weighted one (0.74). This difference can be attributed to the underestimation of mortality by the unweighted model (SN: 0.55 and SP: 0.80), better reflecting the low prevalence of mortality in the study cohort. In contrast, the weighted model achieves a balanced predictive performance, with sensitivity at 0.67 and specificity at 0.75.

These results present a competitive performance of the simplified hXG-SAPS, on par with widely used scoring systems, despite using limited information. One example is the high AUROC of the weighted simplified hXG-SAPS, which is comparable to

OASIS and APS-III, and of the unweighted version, which is comparable to SOFA and SIRS, as shown in Figure 5.3. Another example is the high O/E ratio of both simplified hXG-SAPS compared to OASIS, SAPS-II, and SOFA. While these results do not extend to accuracy, sensitivity and specificity, the simplified hXG-SAPS is a robust model to discriminate and prioritize patients with only 5 inputs as compared to OASIS (10 features), SAPS-II (15 features) and APS-III (20 features).

5.3.4 hXG-SAPS performance across age and ethnicity stratified cohorts

In this section, we examine the performance of our model within the age and ethnicity subgroups of our dataset as presented in Section 3.4. The results breakdown is provided in Table 5.3. We start with an assessment of the different age groups and finish with an examination of ethnic groups.

5.3.4.1 hXG-SAPS results for age stratified cohorts

A first observation is the consistent performance of the weighted and unweighted hXG-SAPS across age groups, as detailed in Table 5.4. The only exception occurs in the *18-24* years cohort, where the unweighted version outperforms the weighted one in AUROC (0.88 and 0.86) and O/E (0.77 and 0.73).

Method	18-24			25-44			45-64			65-84			Over 85		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
SAPS-II	**0.83**	0.88	**0.83**	0.77	**0.84**	0.77	0.74	0.76	0.73	**0.76**	0.68	0.77	**0.74**	0.62	**0.77**
OASIS	0.75	0.85	0.75	0.77	0.71	0.77	0.69	0.74	0.69	0.69	**0.69**	0.69	0.66	**0.76**	0.63
APS-III	0.82	0.90	0.82	**0.82**	0.77	0.82	**0.77**	0.72	**0.77**	**0.75**	0.66	0.77	0.72	**0.71**	0.72
SOFA	0.73	0.85	0.73	**0.82**	0.78	0.82	0.75	0.63	**0.76**	0.70	0.56	0.72	0.71	0.52	0.75
SIRS	0.61	0.81	0.60	0.58	0.64	0.58	0.61	0.62	0.61	0.52	**0.69**	0.50	0.44	0.72	0.38
XG-SAPS	**0.85**	**0.94**	**0.85**	**0.78**	0.78	**0.78**	**0.78**	**0.78**	**0.78**	**0.76**	**0.72**	0.77	**0.74**	0.63	0.76
OASIS+	**0.87**	**0.99**	**0.87**	0.75	**0.83**	0.74	0.70	**0.78**	0.69	0.69	**0.75**	0.68	0.71	**0.78**	0.70
hXG-SAPS (w)	0.80	**0.93**	0.80	**0.82**	**0.85**	**0.81**	**0.76**	**0.81**	0.75	**0.77**	**0.69**	**0.79**	**0.76**	0.59	**0.79**
hXG-SAPS (u)	**0.83**	0.91	**0.83**	**0.82**	**0.84**	**0.82**	**0.77**	**0.80**	**0.76**	**0.77**	**0.69**	**0.78**	**0.75**	0.60	**0.79**
(S) hXG-SAPS (w)	0.62	0.92	0.61	**0.86**	0.81	**0.86**	0.74	0.75	0.74	0.73	0.67	0.74	0.73	0.62	0.75
(S) hXG-SAPS (u)	**0.83**	0.64	**0.83**	0.77	0.78	0.77	0.75	0.68	**0.76**	**0.77**	0.50	**0.81**	0.73	0.45	**0.80**

Table 5.5: Mean accuracy, sensitivity and specificity for predictive models applied to age cohorts. In bold we highlight the top 3 best performing models for each cohort. Figure C.2 in Appendix C shows the results including the confidence interval.

Despite these minor variations, the weighted and unweighted hXG-SAPS consistently rank within the top three models in terms of AUROC, except in the *18-24* and *Over 85* age groups. In both cases, hXG-SAPS ranks close behind OASIS+, XG-SAPS, and APS-III. Of interest is the outstanding AUROC of OASIS+ (0.96 for *18-24* and 0.79 for *Over 85*). However, a closer examination reveals that OASIS+ tends to

underestimate mortality in the younger cohort and overestimate it in the older one. Surprisingly, the weighted simplified hXG-SAPS matches the performance of widely used scoring systems across cohorts. For example, it performs as well as OASIS for the *0-24* age group, better than APS-III for age groups spanning 25 to 84 years old, and as SAPS-II for age cohorts with patients over 25 years old.

Table 5.5 shows the consistent and high accuracy of both versions of hXG-SAPS across all age cohorts. Notably, both versions achieve the highest accuracy for the 65-84 (0.77 for both) and Over 85 (0.76 and 0.75) groups and the second best for the 45-64 (0.76 and 0.77), closely behind XG-SAPS (0.78). Interestingly, we observe an exceptional performance of OASIS+ and XG-SAPS (0.87 and 0.85, respectively), relegating unweighted hXG-SAPS to the third highest accuracy (0.83) in the 0-24 age group. A closer look reveals that OASIS+ and XG-SAPS show an underestimation of mortality for these young patients, which can explain their high accuracy. Additionally, Table 5.5 shows a consistent and high O/E performance of hXG-SAPS and the simplified hXG-SAPS across age groups, while all other models show a noticeable poor. These results point to the robustness of our approach even within distinct age subgroups in our study cohort.

While these results confirm the previously reported trend of increasing mortality overestimation for older patients, it is noteworthy that hXG-SAPS displays a more balanced performance across all cohorts in terms of AUROC and accuracy. This is evident when comparing with OASIS+ and XG-SAPS which show high AUROC and accuracy for younger cohorts at the cost of underestimating mortality.

5.3.4.2 hXG-SAPS results for ethnicity stratified cohorts

Table 5.6 reveals a consistent and high AUROC of hXG-SAPS across all ethnicity cohorts, regardless of its version. The weighted and unweighted hXG-SAPS show the highest AUROC in the *Black* (0.88 and 0.87) and *Other* (0.82 and 0.82) ethnicity cohorts. While it may not top the performance in other cohorts, hXG-SAPS remains remarkably close to the best-performing methods. For example in the White group, it ranks second-highest AUROC (0.83), just behind OASIS+ (0.90) and on par with XG-SAPS (0.83). In the Asian and Hispanic cohorts, its performance is significantly close to the top-performing methods despite ranking third and fourth respectively. Notably, the simplified hXG-SAPS display AUROC results comparable to widely used methods. For example, the weighted version (0.78) surpasses OASIS (0.76) and SOFA (0.71) in the White group. Another example is the unweighted version (0.65) outperforming

Method	White		Black		Asian		Hispanic		Other	
	AUROC	O/E	AUROC	O/E	AUROC	O/E	AUROC	O/E	AUROC	O/E
SAPS-II	**0.80**	0.34	**0.86**	0.26	**0.84**	0.27	**0.79**	0.26	**0.76**	0.50
OASIS	0.76	0.23	0.78	0.17	0.80	0.17	0.70	0.16	0.74	0.33
APS-III	0.78	0.46	0.81	0.32	0.86	0.36	0.66	0.33	**0.77**	**0.69**
SOFA	0.71	0.48	0.75	0.35	0.82	0.40	0.63	0.34	0.68	**0.66**
SIRS	0.61	0.15	0.65	0.11	0.61	0.11	0.57	0.10	0.60	0.22
XG-SAPS	**0.83**	**0.97**	**0.88**	**0.71**	**0.87**	0.71	**0.78**	**0.80**	**0.82**	1.36
OASIS+	**0.90**	0.70	0.83	0.53	0.80	0.59	**0.81**	0.56	0.78	**1.04**
hXG-SAPS (w)	**0.83**	**1.00**	**0.88**	**0.69**	**0.84**	**0.73**	0.77	**0.83**	**0.82**	1.39
hXG-SAPS (u)	**0.83**	**0.99**	**0.87**	**0.71**	**0.83**	0.72	0.76	**0.83**	**0.82**	1.39
(S) hXG-SAPS (w)	0.78	**1.00**	0.76	0.63	0.79	**0.78**	0.66	0.68	**0.77**	1.45
(S) hXG-SAPS (u)	0.71	**0.99**	0.69	0.63	0.64	**0.77**	0.65	**0.69**	0.68	1.49

Table 5.6: AUROC and Observed-to-Expected ratio summary of all predictive models applied to ethnicity cohorts. In bold we highlight the top 3 best performing models for each evaluation metric in each cohort.

SOFA (0.63) and on par with APS-II (0.66) in the Hispanic cohort.

These positive results extend to accuracy, specificity and sensitivity as evidenced in Table 5.7. The weighted hXG-SAPS exhibits the best performance for the *White* (0.77) and *Black* (0.81) cohorts, on par with XG-SAPS. For the *Asian* cohort, APS-III appears as the top-performing model in terms of accuracy, sensitivity and specificity (ACC: 0.87, SN: 0.83, SP: 0.88), followed by the unweighted hXG-SAPS (ACC: 0.81, SN: 0.78, SP: 0.81). Interestingly there is a noticeable lack of accuracy for the *Hispanic*, and *Other* cohorts across all predictive models with a general tendency to overestimate mortality as indicated by all models showing a significantly higher specificity than sensitivity. But, maybe more interesting is that the simplified unweighted hXG-SAPS displays the highest accuracy for these cohorts (0.72 for *Hispanic* and *Other*). This suggests that the use of hSBM is linked to an increase in accuracy and mitigation of mortality overestimation in these cases.

Similarly to the case for age cohorts, both hXG-SAPS and XG-SAPS outperform all other severity scores in terms of calibration, as measured by the Observed-to-Expected mortality ratio presented in Table 5.7. In the White cohort, hXG- SAPS achieves the highest calibration (1.00), and the same trend is observed for the Hispanic cohort (0.83). A very interesting example is that of the Asian cohort, where the simplified versions of XG-SAPS outperform XG-SAPS as the best model (0.78 and 0.73). In the case of Black and Other cohorts, no significant improvement is noted. Notably, hXG-SAPS and its simplified version demonstrate either better or equivalent results compared to XG-SAPS, OASIS+ and the other scoring systems.

Method	White			Black			Asian			Hispanic			Other		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
SAPS-II	**0.76**	0.71	**0.76**	**0.81**	0.81	**0.81**	**0.87**	0.69	**0.89**	**0.67**	0.86	**0.66**	0.70	0.74	0.69
OASIS	0.70	**0.73**	0.70	0.73	0.74	0.73	0.71	**0.84**	0.70	0.61	0.77	0.59	0.69	0.69	0.69
APS-III	**0.75**	0.71	**0.76**	**0.76**	0.79	**0.76**	**0.87**	**0.83**	**0.88**	**0.65**	0.68	0.64	0.73	0.71	0.73
SOFA	0.73	0.60	0.75	0.70	0.73	0.70	0.69	**0.84**	0.67	0.58	0.68	0.57	0.67	0.62	0.67
SIRS	0.52	0.67	0.50	0.52	0.74	0.50	0.68	0.57	0.69	0.54	0.60	0.54	0.64	0.51	0.66
XG-SAPS	**0.77**	**0.75**	**0.78**	**0.81**	**0.84**	**0.81**	0.76	**0.92**	0.75	0.63	**0.93**	0.61	**0.73**	**0.80**	**0.72**
OASIS+	0.71	**0.78**	0.70	**0.80**	0.77	**0.80**	0.72	**0.83**	0.71	**0.72**	**0.85**	**0.71**	**0.72**	0.74	**0.71**
hXG-SAPS (w)	**0.77**	**0.75**	**0.77**	**0.81**	**0.86**	**0.80**	0.78	0.82	0.78	0.63	**0.90**	0.61	**0.71**	**0.82**	0.69
hXG-SAPS (u)	**0.77**	**0.75**	**0.77**	**0.80**	**0.85**	**0.80**	**0.81**	0.78	**0.81**	0.64	**0.87**	0.62	**0.71**	**0.81**	**0.70**
(S) hXG-SAPS (w)	0.74	0.69	0.75	0.71	0.73	0.71	0.74	**0.83**	0.74	**0.72**	0.67	**0.73**	**0.71**	0.73	**0.70**
(S) hXG-SAPS (u)	**0.77**	0.58	**0.79**	0.72	0.63	0.73	0.73	0.60	0.74	**0.72**	0.56	**0.73**	**0.72**	0.53	0.76

Table 5.7: Mean accuracy, sensitivity and specificity for predictive models applied to age cohorts. In bold we highlight the top 3 best performing models for each cohort. Figure C.3 in Appendix C shows the results including the confidence interval.

5.4 Discussion

Building upon our networked-based patient model and applying community detection, we developed a hierarchical mortality predictive model that overcomes several of the drawbacks of current methods. Our model can handle incomplete ICU data, incorporate patient heterogeneity and automatically show calibrated outcomes. To our knowledge, this is the first work introducing the use of community detection to build a mortality-predictive model for ICU patients. Our model presents several advantages, which are outlined below.

State-of-the-art performance: Our model exhibits state-of-the-art performance across various metrics, significantly outperforming widely used severity of scores and on par with machine learning models. Notably, hXG-SAPS achieves a high AUROC while striking an excellent balance between true positives and negatives, as demonstrated by its sensitivity and specificity. Furthermore, the simplified hXG-SAPS shows a discriminatory power comparable to widely utilised scores such as OASIS, and APS-III in its weighted version, and to SOFA and SIRS in its unweighted version. Additionally, as shown in Section 5.2.4, our model shows an excellent output calibration as evidenced by its Observed-to-Expected mortality ratio.

Enhanced heterogeneity handling: The performance of our model extends to the various age and ethnicity population cohorts described in Section 5.3.4.1. We find that hXG-SAPS consistently demonstrates the highest discriminant power, accuracy and sensitivity/specificity trade-off across these cohorts. Prominent examples can be observed in the older age cohorts - 65-84 and over 85 years old - and the White and Black ethnicity cohorts. In these instances, hXG-SAPS show the highest AUROC and accuracy, while maintaining top sensitivity/specificity trade-offs. Surprisingly, even with

a decreased performance, the simplified hXG-SAPS consistently outperforms SOFA and SIRS, ranking amongst the top three scoring systems.

Flexibility to varying levels of available data: The hierarchical organization of features yields a predictive model that combines two models simultaneously: one trained with all block features and another with reduced block clusters. This results in a simple yet comprehensive model for patient assessment, capable of adapting its outputs to the available information. In addition to the previously discussed performance of the weighted hXG-SAPS, our model can adapt to even less information by means of the simplified weighted hXG-SAPS to provide predictions comparable to widely used severity scores such as SOFA and SIRS and O/E ratio comparable to the XG-SAPS. Furthermore, although losing calibration at a cohort level, the unweighted hXG-SAPS maintains a favourable O/E ratio at a sample level, considering only five binary features.

Automatic model simplification: Our findings underscore our model's ability to automatically adapt to the training data, capturing relevant patterns from the clinical variables and simplifying the use of our model. Notably, this is achieved without altering the SAPS-II input features, ensuring consistency with this well-established model and transparency to the user. This is made possible due to the identification of blocks of features that are statistically identical in their connectivity behaviour to patients. We use these blocks to simplify the model, for example by combining categories (e.g. patients between 69 and 84 years old). Since nodes within blocks present similar connectivity patterns (Peixoto, 2014b), our simplified models preserve the underlying relationship between feature nodes and patients. This is evidenced by the nearly identical performance of the weighted and unweighted hXG-SAPS and SAPS presented in previous sections.

Robustness to missing data and automatic imputation: A notable characteristic of our data modelling is that our models do not require imputation; instead, no connection to a block is registered if the value is missing. Moreover, by considering only the patient's links to the feature blocks, we omit the subscores, limiting the connection between our model and SAPS-II only to the clinical variables and their ranges.

Automatic output calibration: hXG-SAPS shows excellent calibration in the entire test dataset and across different patient cohorts, consistently achieving top O/E ratio. Notably, this is achieved automatically, without explicitly accounting for patient cohorts. This is in contrast to models such as in SAPS-II, specifically calibrated for different ethnicity cohorts (Moreno et al., 2005). Clear results are observed in

the younger and Hispanic cohorts, where hXG-SAPS significantly outperforms XG-SAPS. Furthermore, the simplified hXG-SAPS maintains this level of performance across most cohorts. These results point to our approach's ability to accommodate the heterogeneity of ICU patients.

Limitations and Future Work

While our results emphasize the strengths of our approach, it is important to acknowledge some limitations to our findings. In the following, we outline some of these and discuss opportunities for further research to enhance the robustness of our approach.

One limitation to the wider validity of our findings is our reliance on data from a single source. This strongly correlates our findings to the specific patient population, hospital protocols, and resources used in the creation of MIMIC-III. To address this, we plan to expand our research to include other data sources. By doing this, we aim to validate the methodological advantages of our approach across different data sources.

Another limitation in our current results is the reliance on clinical variables and subscores defined in SAPS-II. The use of a SAPS-II links our findings to the quality of this particular score, making it hard to determine the extent to which our results are influenced by this underlying system. To mitigate this, we employ OASIS+ and XG-SAPS as benchmarks to better understand the impact of our methodology.

A final limitation relates to the limited inclusion of medical knowledge into the predictive model. We mitigate this by aligning our model with medical knowledge via the use of SAPS-II, a well-established and medically validated score. Regardless, the inclusion of domain knowledge such as mortality prevalence or feature importance in the prediction process still constitutes an open research area.

An interesting research opportunity comes from leveraging our hSBM models and network reconstruction techniques to replace machine learning. In particular, the use of link prediction for mortality prediction holds the potential to address the limitations just discussed. Initially, it is possible to lean on the scalability of link prediction to model the patient and clinical variables directly, eliminating the reliance on SAPS-II subscores. Furthermore, recent work offers techniques to include domain knowledge, in the form of informative priors, into the link prediction process (Martin et al., 2016). By doing so, we could include valuable domain insights, such as the expected mortality prevalence, enhancing the ability of our approach to bridge the gap between medical and data-driven knowledge.

5.5 Conclusion

The importance of patient severity assessment in the ICU and the wealth of available data has fostered research into models supporting decision-making in the ICU. Notably, severity of illness scores and machine learning models built upon these scores constitute useful tools to aid medical staff in the ICU. However, the challenges of incomplete data, the need for imputation, and patient heterogeneity affect their performance. We address these issues by creating a network-enhanced prediction model. Our model achieves consistent performance across different patient cohorts, avoiding imputation despite missing data. Additionally, it offers flexibility to different levels of available information and does not require outcome calibration.

Besides the specific contributions in the task of mortality risk assessment, our findings contribute to answering our research questions, outlined in Chapter 1:

- **RQ1**: Our model can automatically impute missing data without making assumptions about why the data is missing. Additionally, the hierarchical nature of our approach allows our model to provide valuable support under different levels of data availability. All of this points to the better suitability of our approach to the ICU data than the benchmark methods studied.

- **RQ2**: The excellent performance of our model across patients' age and ethnicity highlights the ability of our approach to simultaneously incorporate patterns in the characteristics of patients into the clustering of features. As discussed in 2.3.3, this is due to the hierarchical nature of the hSBM, allowing it to incorporate important aspects of the diversity of ICU patients into the model.

- **RQ3**: Using the hierarchical clustering model, we automatically extract relevant relationships between clinical variables and patients. Capturing this information in the data-block matrix, we include this information in our predictive model. This reduces the necessity for human intervention in fitting statistical or machine learning models, such as assigning subscore values. In this way, our approach simplifies the use of our model in the ICU.

In conclusion, our research presents a novel use of network sciences for the support of patient mortality assessment addressing the ICU-specific challenges discussed in Section 1. These results provide strong evidence for the potential of network analysis and community detection in improving predictive models for ICU patient outcomes.

In the next chapter, we take a step back to summarize the contributions of our current research and general research avenues that might extend our contributions.

Chapter 6

Conclusions

In this concluding chapter, we take a step back and consider the broader benefits of our approach in the ICU context. To this end, we revisit our findings and their alignment with our research questions outlined in Chapter 1, reflect on the contributions and limitations of our work, and outline directions for future research.

6.1 Our contributions

In this section, we outline the contributions of our work. We start by presenting how our approach addresses our research questions and then move to highlight some unexpected but valuable observations made during our work.

6.1.1 Addressing our research questions

We have been able to achieve various benefits across tasks while addressing the challenges presented in our research question. We reflect on these next.

RQ1: How can we improve the robustness of machine learning models over missing and sparse ICU data?

Our results show how the use of network-based representation of patients and community detection offers a promising answer to this question. Firstly, the rich network representation of patients and the non-parametric community detection approach to unveil statistically robust structures in the network palliates the effects of missing and incorrect data. Additionally, the non-parametric and hierarchical aspects of our approach enable our models to capture relevant patterns even with sparse data.

Two notable examples of the resilience of our approach to incomplete and missing data are presented in Chapter 5. Firstly, our scoring system can automatically impute missing data without assumptions about the reasons for the missing data. Secondly, our approach exhibits effective handling of unbalanced data, as shown by its top performance in predicting mortality, highly unbalanced in our cohort study, without data prepossessing or model calibration. Additionally, Chapter 4 demonstrates our approach's ability to handle sparse data. Our findings show that our clustering models effectively incorporate the medically relevant yet sparse comorbidities, along with the denser demographic and admission-type data to identify multimorbidity profiles. This is not observed in benchmark methods, which mostly rely on demographic information and admission type to group patients.

In summary, our approach effectively addresses key challenging aspects of ICU data, including sparsity, unbalance, and incompleteness of data, which traditional methods struggle to handle. This resilience ensures that our models remain reliable and accurate when dealing with ICU data.

RQ2: How can we enhance machine learning models to incorporate the heterogeneity of ICU patients?

Our approach finds in the use of network science and non-parametric approaches a promising answer to this question. Firstly, our network-based representation allows the explicit encoding of complex non-linear relationships between patients and clinical variables, which traditional approaches may struggle to represent. Furthermore, the use of non-parametric community detection allows models to adapt to various patient groups, accommodating diverse patient profiles.

Chapter 5 serves to illustrate our approach's ability to handle patient heterogeneity. There, it demonstrates top performance in mortality prediction across the entire sample as well as across age and ethnicity cohorts in terms of ROC-AUC, accuracy, and outcome calibration. This points to our model's ability to identify patterns specific to similar groups of patients. Notably, our approach can identify patient groups in aspects not explicitly represented in the dataset, as shown in Chapter 4. In this case, our model can distinguish patient groups not solely based on specific demographics or morbidities but also by considering multimorbidity counts. Our approach is able to capture this reportedly relevant aspect of the data, noticeably missing from the benchmark methods. By considering multimorbidity count, our model uncovers patient groups that

other methods overlook, such as those without comorbidities.

In summary, our approach is able to handle the heterogeneity of patients by accurately identifying groups of homogeneous patients. This allows our models to naturally accommodate the diverse range of ICU patients, ultimately leading to more precise and patient-specific outcomes.

RQ3: How can we leverage available ICU data to simplify the training and validation of machine learning models?

The use of a non-parametric approach represents an answer to this research question in our approach. This aspect significantly simplifies model training and selection, offering a data-driven solution to this technically demanding task. This is evident in different aspects of our work.

In Chapter 4, this is illustrated as all the parameters of the clustering model are automatically inferred, eliminating the need for manual intervention. This is in contrast to the benchmark methods that require significant effort to determine the optimal cluster numbers and extensive model validation. Similarly, in Chapter 5, the use of hSBM automatically creates a mapping of features into a simplified input space, sequentially used for mortality prediction. Another important aspect of simplification relates to data handling. Unlike the benchmark methods in Chapter 4, our approach does not rely on assumptions regarding feature distributions or independence. Furthermore, it eliminates the necessity for explicit data imputation before model training, in contrast with some of the benchmark methods in Chapter 5.

In conclusion, our approach effectively simplifies the training and validations of models for the ICU, lifting parts of the technical challenges related to the application of machine learning models in the ICU context.

6.1.2 Other benefits of our approach

Our approach presents unexpected but valuable elements contributing to the usability of our models in the ICU context. One such element of usability stems from our approach's robustness to ICU data. This robustness and adaptability to data quality and availability offer flexibility to clinicians as they can use our models even when data is not optimal. This is particularly clear in Chapter 5, as our model exhibits flexibility and adaptability to varying data availability levels. As a result, our model enables patient assessments even when limited clinical variables are unavailable.

6.2 Future directions

As presented in the previous section, our approach's adaptability, robustness, and simplified model selection, coupled with enhanced usability, offer notable benefits for decision-making support in the ICU. Beyond these advantages, our work also opens exciting research opportunities, offering solutions to current limitations and enhancing our contributions. In the following, we highlight three immediate extensions to our work, as concrete examples of future research and to illustrate how these would tackle some of the current limitations of our work.

Enrich clinical data representation: Managing extensive clinical data presents both opportunities and challenges. Current approaches typically preprocess data to reduce the input space dimensionality, as illustrated by multimorbidity indices in Chapter 4 and severity of scores in Chapter 5. While useful, this strategy presents a limitation, potentially losing important information during the dimensionality reduction. Leveraging network science's scalability and resilience to sparse data we can directly study the available clinical data. As a first immediate step, we plan to extend our multimorbidity profiles study using ICD-9 instead of multimorbidity indices.

Network reconstruction for mortality prediction: Community detection models, as shown in Chapters 4 and 5, are powerful for understanding underlying clustering structures. However, these models are also probabilistic with additional potential applications. For example, they can be used for network reconstruction, with applications to outcomes prediction and data imputation. An immediate research direction is to extend our work in Chapter 5, replacing the use of machine learning models for mortality prediction by link prediction via network reconstruction leveraging our existing clustering models.

Incorporation of medical knowledge: Up to this point, we have explored the benefits of simplifying the application of machine learning in the ICU context. By doing this, we aimed to lift some of the technical complexities to bridge machine learning and the medical domain. Nonetheless, the data-driven nature of our approach presents limitations, as it does not fully incorporate medical knowledge into our models. A step further would be to better include medical knowledge in our approach. This can be achieved in various ways by leveraging our network-based representation and Bayesian inference process. For example, by adding weights to our patients-morbidity network

we could point to relationships that are medically more relevant for mortality or sepsis. Another option and an immediate research direction is the inclusion of priors to convey mortality expectations for in-hospital death prediction via network reconstruction.

6.3 Concluding remarks

In this thesis, we have introduced an approach to uncover underlying structures in data that can help the understanding of complex relationships between patients, clinical variables and medical outcomes. From the beginning, we envisioned a solution that could account for the complexities of the ICU domain in terms of the clinical data and the intricacies of the medical processes, tackling the issues that widely used methods struggle with.

In pursuit of this vision, we have realized the advantages of combining network-based representation of clinical data, network science and machine learning methods. More specifically, we successfully applied this approach to identify multimorbidity profiles for ICU patients in Chapter 4 and construct a hierarchical mortality prediction model in Chapter 5. In contrast to current methods, our approach creates models that are able to automatically handle missing and sparse data, account for the heterogeneity of the ICU population, and eliminate human intervention in the model selection process. Our work not only offers innovative solutions but also opens new research directions to further enhance our contributions.

To conclude, this thesis represents a step forward in the use of networks and machine learning methods to effectively support decision-making in the ICU in its whole complexity. This recognizes the complexities of the clinical data and importantly opens research avenues to deepen the integration of network science, machine learning and medical knowledge, all in service of enhancing patient care in critical situations.

Appendix A

Background

A.1 Severity of illness score tables

In this section we complement the information provided for the severity of illness scores introduced in Section 2.1.4 and later utilized throughot Chapter 5.

- **OASIS**: This scoring system is presented in Figure 2.1. The mortality probability function (MP) part of this scoring system:

$$MP = 1/(1+\exp(-(-6.1746+0.1275\cdot \text{OASIS SCORE}))) \tag{A.1}$$

- **SOFA**: The detail scoring system is provided in table A.1. There is no mortality probability function as part of this severity score.

- **SIRS**: The detail scoring system is provided in table A.2. There is no mortality probability function as part of this severity score.

- **SAPS-II**: This scoring system is presented in Figure 5.1. The mortality probability function (MP) part of this scoring system:

$$MP = 1/(1+exp(-(-7.7631+0.0737*\text{sapsii}+0.9971*(ln\text{sapsii + 1})))) \tag{A.2}$$

- **APS-III**: The detail scoring system is provided in table A.1. The mortality probability function (MP) part of this scoring system:

$$MP = 1/(1+exp(-(-4.4360+0.04726*\text{APS-III SCORE}))) \tag{A.3}$$

	SOFA subscores			
	1	2	3	4
Respiratory PaO_2/FiO_2 (mmHg)	< 400	< 300	< 200	< 100
Coagulation Platelets x $10^3/mm^3$	< 150	< 100	< 50	< 20
Liver Bilirubin (mg/dL)	1.2 - 1.9	2.0 - 5.9	6.0 - 11.9	> 12.0
Cardiovascular Hypertension	MAP < 70 mmHg	Dopamine < 5 or Epinephine > 0.0 or Norepinephrin > 0.0	Dopamine > 5 or Epinephine < 0.1 or Norepinephrin < 0.1	Dopamine > 15 or Epinephine > 0.1 or Norepinephrin > 0.1
Central Nervous System Glasgow Coma Scale	15	13 - 14	10 - 12	6 - 9
Renal Creatinine (mg/dL)	< 1.2	1.3 - 1.9	2.0 - 3.4	> 3.4

Table A.1: SOFA subscores and clinical ranges. MAP stands for Mean arterial pressure

	SIRS subscores	
	0	1
Body temperature (ºC)	between 36 and 38	< 36 or > 38
Heart rate (beats/min)	≤ 90	> 90
White blood cell count (10^3/dL)	between 4 and 12	< 4 or > 12
Respiratory rate (breaths/min)	≤ 20	> 20

Table A.2: SIRS subscores and clinical ranges

Score	20	16	13	8	2	0	Body temperature	4			
Range	< 32.9	33-33.4	33.5-33.9	34-34.9	35-35.9	36-39.9	(°C)	> 40			
				8	5	0	Heart beat	1	5	7	13	17
				< 39	40-49	50-99	(beats/minute)	100-109	110-119	120-139	140-154	> 155
		23	15	7	6	0	Mean BP	4	7	9	10	
		< 39	40-59	60-69	70-79	80-99	(mmHg)	100-119	120-129	130-139	> 140	
			17	8	7	0	Respiratory rate	6	9	11	18	
			> 5	6-11	12-13	14-24	(breaths/min)	25-34	35-39	40-49	> 50	
			15	5	2	0	PaO2 mm Hg					
			< 49	50-69	70-79	> 80						
						0	A-aDO_2 mm Hg	7	9	11	14	
						< 100		100-249	250-349	350-499	> 500	
					3	0	Hematrocrit (%)	3				
					> 40.9	41-49		> 50				
				19	5	0	White blood cell	1	5			
				< 1.0	1.0-2.9	3.0-19.9	(count x 10^9/L)	20-24.9	> 25			
					3	0	Creatinine	4	7	With acute renal		
					< 0.4	0.5-1.4	(mg/dL)	1.5-1.94	> 1.95	failure		
						0	Creatinine	10	Without acute renal			
						0-1.4	(mg/dL)	> 1.5	failure			
	15	8	7	5	4	0	Urine output	1				
	< 399	400-599	600-899	900-1499	1500-1999	2000-3999	(ml/24 hours)	> 4000				
						0	BUN	2	3	11	12	
						< 16.9	(mg/dl)	17-19	20-39	40-79	> 80	
				3	2	0	Sodium	4				
				< 119	120-134	135-154	mM/L	155				
				11	6	0	Albumina	4				
				> 1.9	2.0-2.4	2.5-4.4	(g/dl)	> 4.5				
						0	Billirubin	5	6	8	16	
						< 1.9	(mg/dl)	2.0-2.9	3.0-4.9	5.0-7.9	> 8.0	
				8	9	0	Glucose	3	5			
				< 39	40-59	60-199	(mg/dl)	200-349	> 350			

Table A.3: APS-III subscores and clinical ranges

Appendix B

Improving identification of multimorbidity profiles for ICU patients

B.1 Charlson clustering models stability

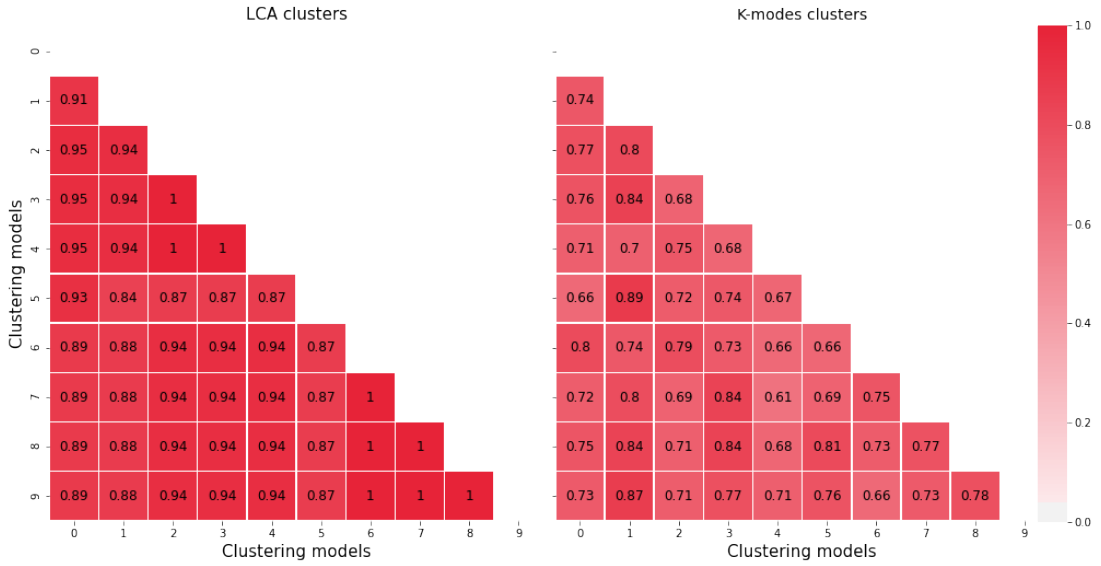

Figure B.1: Comparing Mutual Information (MI) in the Top 10 Charlson multimorbidity profile models generated by LCA (left) and K-modes (right). Models are labelled from 0 to 9 based on lower entropy, with 0 being the best and 9 the least optimal. Notably, MI is considerably greater in LCA partitions compared to K-modes partitions. The optimal model for each method (model 0) maintains the highest MI with other models, emphasizing its consistency across all models.

B.2 Charlson clusters outcomes prevalence

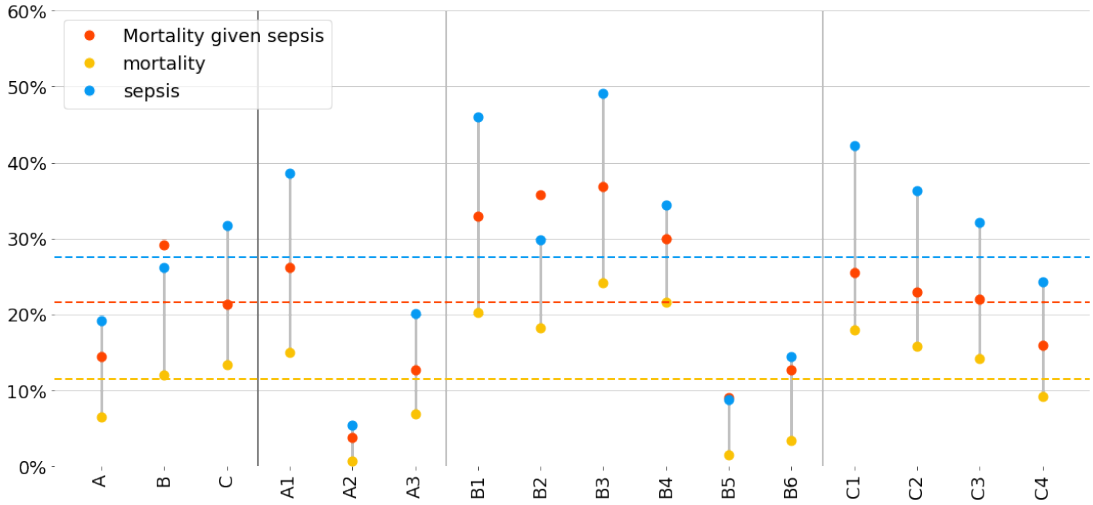

Figure B.2: Mortality, sepsis, and mortality-given-sepsis for Charlson hSBM clusters.

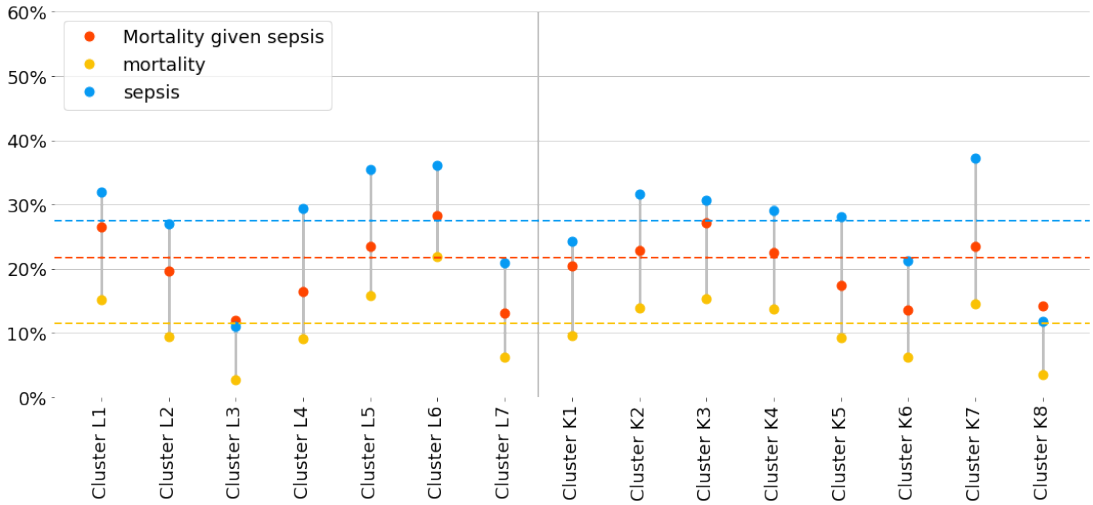

Figure B.3: Sepsis, mortality, and mortality given sepsis across Charlson LCA (left) and K-modes (right) clusters.

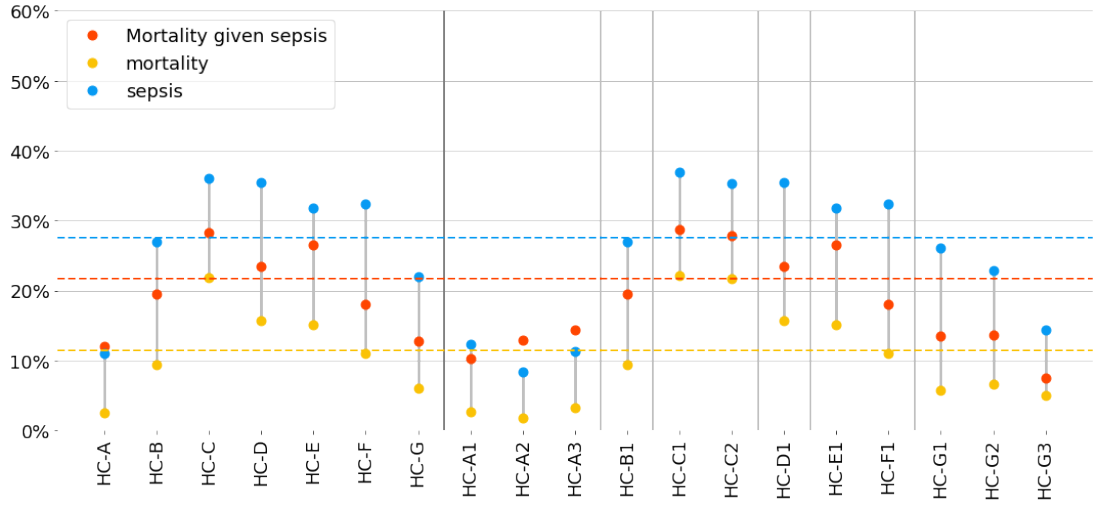

Figure B.4: Sepsis, mortality, and mortality given sepsis across Charlson agglomerative clustering, 7 clusters (left) and 12 clusters (right).

B.3 Charlson clusters comorbidity count

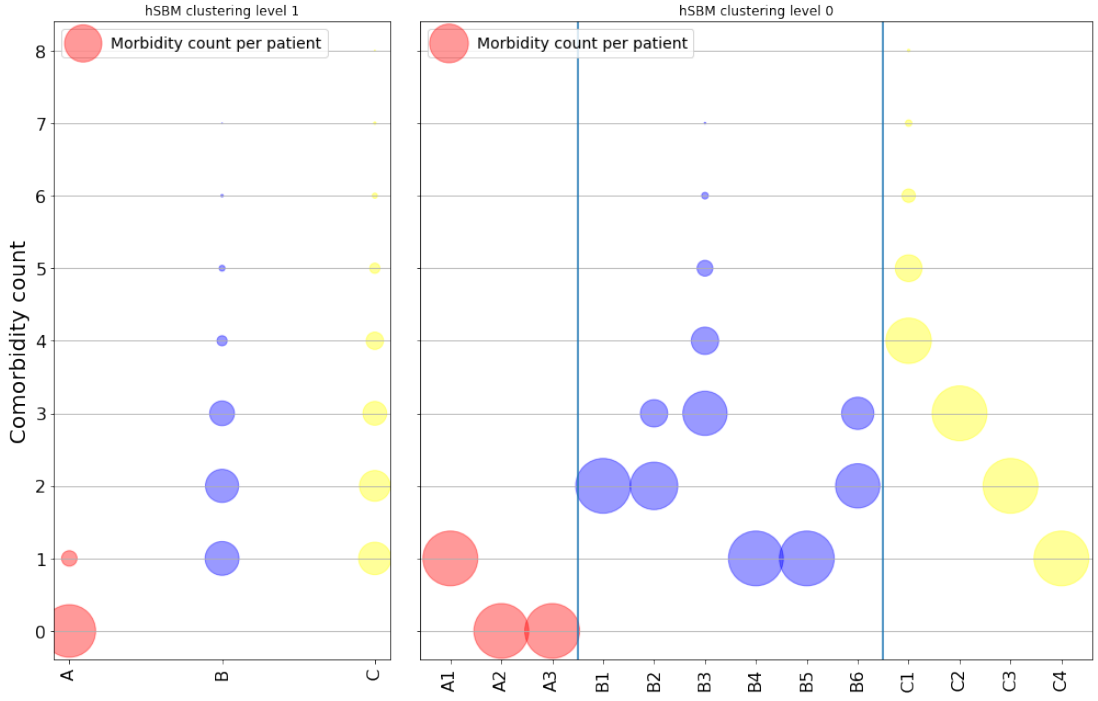

Figure B.5: Count of Charlson morbidities per patient, at both the intermediate (left) and bottom (right) hierarchical levels. Circle size represents the number of patients with a specific count of morbidities.

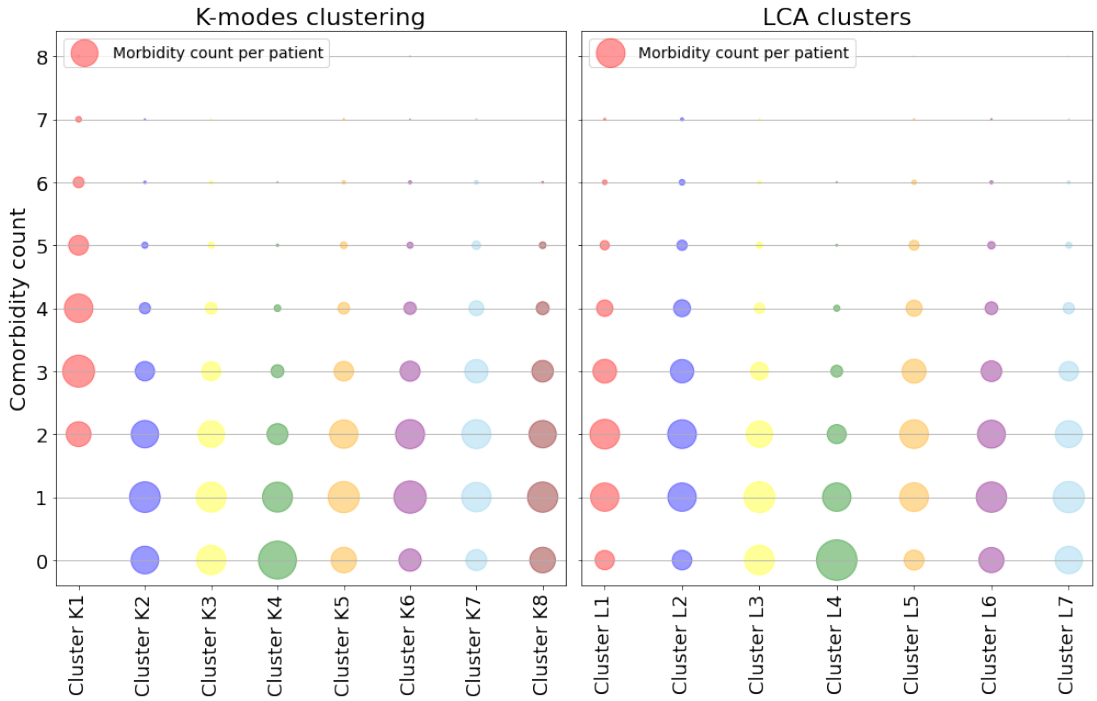

Figure B.6: Count of Charlson-defined comorbidities per patient for LCA (right) and K-modes (left) clusters. Circle size represents the frequency of patients with a specific number of morbidities.

B.4 hSBM clustering similarities across multimorbidity indices

Patients over 45 with heterogeneous multimorbidity count (Elixhauser clusters F1$_e$, F2$_e$, F3$_e$, and F4$_e$ & Charlson clusters B4$_c$, C1$_c$, B5$_c$, B6$_c$ and C2$_c$)

The parent clusters Elixhauser F$_e$ and Charlson B$_c$ and C$_c$ show predominantly older patients with multiple co-occurring morbidities, resembling Elixhauser clusters L1$_e$, L3$_e$ K6$_e$ and HC-B$_e$, and Charlson clusters L3$_c$, L4$_c$, K7$_c$ and HD-C$_c$ from the benchmark. But this resemblance disappears when inspecting the ramifications of the inferred hSBM clusters at the bottom hierarchical level. A distinctive feature setting the hSBM clusters apart from the benchmarks is the multimorbidity count, which influences outcomes across different indices.

Elixhauser cluster F1$_e$ and Charlson cluster B4$_c$ encompass patients strictly over 85 years old, with a higher presence of females and a moderately low number of morbidities (between 2 and 4 for F1$_e$ and strictly 1 for B4$_c$). Both exhibit virtually the same rates of mortality (21.6% for both) and mortality given sepsis (30% for both). Despite differences in morbidities, these clusters share a high prevalence of congestive heart failure and a moderately high occurrence of peptic ulcers. Additionally, cluster B4$_c$ shows a very high level of dementia, consistent with the age of the population. Lastly, **Elixhauser cluster F3$_e$ and Charlson clusters B5$_c$ and B6$_c$** are similar, grouping mostly elective patients between 45 and 64 years old with a low number of morbidities (strictly 2 F3$_e$ and strictly 1 for B5$_c$ and 2 or 3 for B6$_c$). The main difference between these clusters and F1$_e$ and B4$_c$ lies in the age of patients, as they do not consider patients over 85 years old. This age gap and the low comorbidity count might explain the lower rates of mortality (8.75% for F1$_e$, 1.47% for B5$_c$, and 3.35% for B6$_c$) and mortality given sepsis (20% for F1$_e$, 8.98% for B5$_c$, and 12.73% for B6$_c$).

The level of detail achieved by the hSBM is not captured by the seemingly similar Elixhauser clusters L5$_e$, L6$_e$, K7$_e$, and HC-B$_e$ or Charlson clusters L3$_c$, L4$_c$, K7$_c$ and HC-C$_c$. While these clusters capture parts of these profiles, such as higher rates of older patients, they fail to capture elements such as the morbidity count, and so the differences between clusters do not translate into differences in medical outcomes.

Older patients with multimorbidity (Elixhause cluster E1$_e$ & Charlson cluster C3$_c$)

These are large clusters comprising older patients, particularly in the 65 to 84 age group and with a moderate number of comorbidities, 3 or 4 for E1$_e$ and 2 for C3$_c$. These clusters share patients with a very low prevalence of substance abuse, AIDS and liver disease, but a slightly higher prevalence of everything else. Similarly to the previous case, the most distinctive difference between these clusters lies in a slightly higher prevalence of elective admissions for E1$_e$ and an absence of it for C3$_c$. However, this difference does not lead to negative outcomes as they present close to average rates of mortality, sepsis and mortality given sepsis. Interestingly such "average" clusters are not directly found in the benchmark. This can be seen by the lack of a benchmark cluster with average levels of the three outcomes considered, namely sepsis, mortality and mortality given sepsis as shown in Figure 4.18.

B.5 hSBM clustering differences across multimorbidity indices

Younger patients with substance abuse and non-elective admissions (Elixhauser Clusters B1$_e$ & B2$_e$)

Clusters B1$_e$ and B2$_e$ exhibit significantly lower rates of elective admissions and substantially higher rates of drug abuse than the dataset average (2 and 2.89 times higher, respectively) and alcohol abuse (22.74% and 21.17% respectively versus an 8.42% average in the dataset). Notably, These clusters of younger patients with a high prevalence of substance abuse are consistent with reports in literature (Zador et al., 2019). The main difference between these two clusters is that all patients in B1 have strictly one morbidity each and in B2$_e$ exactly two, as shown in Figure 4.15. These differences manifest in a higher prevalence of AIDS, psychoses and depression in B2$_e$, and a lower mortality rate given sepsis onset for patients grouped in B1$_e$ (Figure 4.17). Concerning morbidity profiles, these clusters relate to clusters L4$_e$ and K5$_e$, which would include both B1$_e$ and B2$_e$. In both cases hSBM provides more insights about the patients. On one hand K5$_e$ does not correlation with medical outcomes, and on the other L4$_e$ does not allow to distinguish the simpler cases encompass in B1$_e$ (only one morbidity) from the more complicated in B2$_e$.

Elective patients between 45 and 64 years old with cancer (Charlson clusters B2$_c$)

This cluster stands out for its composition of patients with cancer, typically presenting a moderate comorbidity count, typically ranging between 2 and 3. The complexity of patient profiles in this cluster likely contributes to the high mortality rates, 18.24% compared to the sample average of 11.42%. However, what is particularly striking is the remarkably high mortality rate given sepsis, which is 64% higher than the overall population (35.69% compared to the sample average of 21.64%). Interestingly, this cluster is only captured by K-modes cluster K8$_c$ but is related to a lower prevalence of adverse events.

Younger patients with aids and liver disease (Charlson clusters B1$_c$ and B3$_c$)

Patients in B1$_c$ and B3$_c$ show a much lower than average prevalence of elective admissions and a substantially higher prevalence of aids and liver disease, both mild and severe. The main factor that distinguishes these two clusters is that in cluster B1 all patients have exactly two morbidities each, whereas in B3$_c$ they all have three or more, as shown in Figure B.5. These differences likely account for the even higher rates of sepsis onset and mortality given sepsis in B3$_c$ (49.12% and 36.84%) compared to B1 (45.95% and 32.85%). Interestingly, while these morbidity profiles are somehow identified by the benchmarks in clusters HC-B1$_c$, K2$_c$ and L1$_c$ they exhibit a normal or low prevalence of mortality, sepsis and mortality given sepsis, as shown in Figure 4.18.

Appendix C

Improving mortality risk assessment for ICU data and patient heterogeneity

C.1 Performance comparison between predictive models for the entire study cohort

Method	AUROC	Accuracy	Sensitivity	Specificity	RSME	E/O ration
SAPS-II	**0.80**	0.74 (0.74,0.75)	0.73 (0.72,0.73)	0.75 (0.74,0.75)	0.32	0.5
OASIS	0.76	0.70 (0.69,0.70)	0.73 (0.72,0.74)	0.69 (0.69,0.70)	0.30	0.74
APS-III	0.78	**0.75 (0.75,0.75)**	0.71 (0.71,0.72)	**0.76 (0.75,0.76)**	**0.29**	**0.99**
SOFA	0.71	0.71 (0.71,0.72)	0.61 (0.61,0.62)	0.73 (0.72,0.73)	0.32	0.5
SIRS	0.61	0.56 (0.55,0.57)	0.63 (0.62,0.65)	0.55 (0.53,0.56)	0.70	0.15
XG-SAPS	0.83	**0.76 (0.76,0.76)**	**0.75 (0.74,0.75**	**0.77 (0.76,0.77)**	**0.28**	**1.03**
OASIS+	0.80	0.71 **(0.70-0.72)**	**0.77 (0.76-0.78)**	0.70 (0.70-0.71)	**0.29**	0.73
hXG-SAPS (w)	0.83	**0.75 (0.75,0.76)**	**0.74 (0.74,0.75)**	**0.76 (0.75,0.76)**	**0.28**	**1.03**
hXG-SAPS (u)	0.83	**0.75 (0.75,0.75)**	**0.75 (0.75,0.75)**	0.75 (0.75,0.76)	**0.28**	**1.03**
(S) hXG-SAPS (w)	0.77	0.74 (0.74,0.74)	0.67 (0.66,0.67)	0.75 (0.74,0.75)	**0.29**	**1.04**
(S) hXG-SAPS (u)	0.70	**0.77 (0.77,0.77)**	0.55 (0.55,0.55)	**0.80 (0.80,0.80)**	0.30	**1.03**

Figure C.1: Performance summary of all predictive models applied to the whole study cohort. For accuracy, sensitivity, and specificity we report the mean of the bootstrapped evaluation and in parenthesis the 95% confidence interval. In bold we highlight the top three best-performing models for each evaluation metric

Method	18-24			25-44			45-64		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
SAPS-II	**0.83 (0.83,0.84)**	0.88 (0.87,0.88)	**0.83 (0.83,0.83)**	0.77 (0.77,0.78)	**0.84 (0.84,0.85)**	0.77 (0.76,0.77)	0.74 (0.73,0.74)	0.76 (0.75,0.76)	0.73 (0.73,0.74)
OASIS	0.75 (0.74,0.76)	0.85 (0.84,0.86)	0.75 (0.73,0.76)	0.77 (0.76,0.78)	0.71 (0.70,0.72)	0.77 (0.77,0.78)	0.69 (0.68,0.70)	0.74 (0.73,0.75)	0.69 (0.68,0.69)
APS-III	0.82 (0.81,0.82)	0.90 (0.89,0.90)	0.82 (0.81,0.82)	**0.82 (0.81,0.82)**	0.77 (0.76,0.77)	**0.82 (0.81,0.82)**	**0.77 (0.76,0.77)**	0.72 (0.71,0.72)	**0.77 (0.77,0.78)**
SOFA	0.73 (0.73,0.73)	0.85 (0.85,0.86)	0.73 (0.72,0.73)	**0.82 (0.82,0.82)**	0.78 (0.78,0.79)	**0.82 (0.82,0.83)**	0.75 (0.74,0.76)	0.63 (0.62,0.64)	0.76 (0.75,0.77)
SIRS	0.61 (0.60,0.61)	0.81 (0.80,0.81)	0.60 (0.59,0.61)	0.58 (0.57,0.59)	0.64 (0.62,0.65)	0.58 (0.56,0.59)	0.61 (0.60,0.62)	0.62 (0.61,0.63)	0.61 (0.60,0.62)
XG-SAPS	**0.85 (0.84,0.85)**	**0.94 (0.94,0.95)**	**0.85 (0.84,0.85)**	**0.78 (0.78,0.78)**	0.78 (0.78,0.78)	0.78 (0.78,0.79)	**0.78 (0.78,0.78)**	**0.78 (0.78,0.78)**	**0.78 (0.78,0.79)**
OASIS+	**0.87 (0.87,0.87)**	**0.99 (0.99,1.00)**	**0.87 (0.86,0.87)**	0.75 (0.74,0.75)	**0.83 (0.83,0.84)**	0.74 (0.74,0.75)	0.70 (0.70,0.71)	**0.78 (0.77,0.79)**	0.69 (0.69,0.70)
hXG-SAPS (w)	0.80 (0.80,0.81)	**0.93 (0.92,0.93)**	0.80 (0.79,0.80)	**0.82 (0.81,0.82)**	**0.85 (0.84,0.85)**	**0.81 (0.81,0.82)**	**0.76 (0.75,0.76)**	**0.81 (0.81,0.81)**	0.75 (0.75,0.76)
hXG-SAPS (u)	**0.83 (0.82,0.83)**	0.91 (0.91,0.92)	**0.83 (0.82,0.83)**	**0.82 (0.81,0.82)**	**0.84 (0.84,0.85)**	**0.82 (0.81,0.82)**	**0.77 (0.76,0.77)**	**0.80 (0.80,0.81)**	**0.76 (0.76,0.77)**
(S) hXG-SAPS (w)	0.62 (0.61,0.63)	0.92 (0.91,0.93)	0.61 (0.60,0.62)	**0.86 (0.85,0.86)**	0.81 (0.80,0.81)	**0.86 (0.86,0.86)**	0.74 (0.74,0.75)	0.75 (0.74,0.75)	0.74 (0.74,0.75)
(S) hXG-SAPS (u)	**0.83 (0.82,0.83)**	0.64 (0.63,0.65)	**0.83 (0.83,0.84)**	0.77 (0.77,0.77)	0.78 (0.77,0.78)	0.77 (0.77,0.77)	0.75 (0.75,0.75)	0.68 (0.67,0.68)	**0.76 (0.75,0.76)**

Method	65-84			Over 85		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
SAPS-II	**0.76 (0.75,0.76)**	0.68 (0.67,0.68)	0.77 (0.76,0.77)	**0.74 (0.74,0.75)**	0.62 (0.61,0.63)	**0.77 (0.77,0.78)**
OASIS	0.69 (0.68,0.69)	**0.69 (0.68,0.69)**	0.69 (0.68,0.70)	0.66 (0.65,0.66)	**0.76 (0.76,0.77)**	0.63 (0.63,0.64)
APS-III	**0.75 (0.75,0.76)**	0.66 (0.66,0.67)	0.77 (0.76,0.77)	0.72 (0.72,0.72)	**0.71 (0.71,0.72)**	0.72 (0.72,0.73)
SOFA	0.70 (0.69,0.70)	0.56 (0.55,0.57)	0.72 (0.71,0.73)	0.71 (0.71,0.71)	0.52 (0.51,0.52)	0.75 (0.75,0.76)
SIRS	0.52 (0.51,0.53)	**0.69 (0.68,0.70)**	0.50 (0.48,0.51)	0.44 (0.43,0.45)	0.72 (0.71,0.74)	0.38 (0.36,0.40)
XG-SAPS	**0.76 (0.76,0.76)**	**0.72 (0.72,0.72)**	0.77 (0.76,0.77)	**0.74 (0.74,0.74)**	0.63 (0.62,0.64)	0.76 (0.76,0.77)
OASIS+	0.69 (0.69,0.70)	**0.75 (0.75,0.76)**	0.68 (0.68,0.69)	0.71 (0.71,0.71)	**0.78 (0.78,0.79)**	0.70 (0.69,0.70)
hXG-SAPS (w)	**0.77 (0.77,0.78)**	**0.69 (0.68,0.69)**	**0.79 (0.78,0.79)**	**0.76 (0.75,0.76)**	0.59 (0.59,0.60)	**0.79 (0.79,0.80)**
hXG-SAPS (u)	**0.77 (0.76,0.77)**	**0.69 (0.69,0.70)**	**0.78 (0.77,0.78)**	**0.75 (0.75,0.76)**	0.60 (0.60,0.61)	**0.79 (0.78,0.79)**
(S) hXG-SAPS (w)	0.73 (0.72,0.73)	0.67 (0.66,0.67)	0.74 (0.73,0.74)	0.73 (0.72,0.73)	0.62 (0.61,0.63)	0.75 (0.74,0.76)
(S) hXG-SAPS (u)	**0.77 (0.77,0.77)**	0.50 (0.49,0.50)	**0.81 (0.81,0.81)**	0.73 (0.73,0.74)	0.45 (0.45,0.46)	**0.80 (0.79,0.80)**

Figure C.2: Performance summary of all predictive models applied to age stratified cohort. Accuracy, sensitivity, and specificity we report the mean of the bootstrapped evaluation and in parenthesis the 95% confidence interval. In bold we highlight the top three best-performing models for each evaluation metric

Method	White			Black			Asian		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
SAPS-II	**0.76 (0.75,0.76)**	0.71 (0.71,0.72)	0.76 (0.75,0.77)	**0.81 (0.81,0.81)**	0.81 (0.80,0.81)	**0.81 (0.81,0.81)**	**0.87 (0.87,0.88)**	0.69 (0.69,0.70)	**0.89 (0.88,0.90)**
OASIS	0.70 (0.70,0.71)	**0.73 (0.72,0.73)**	0.70 (0.69,0.70)	0.73 (0.72,0.74)	0.74 (0.73,0.74)	0.73 (0.72,0.74)	0.71 (0.70,0.71)	**0.84 (0.83,0.85)**	0.70 (0.69,0.70)
APS-III	**0.75 (0.75,0.76)**	0.71 (0.71,0.72)	0.76 (0.75,0.76)	**0.76 (0.76,0.77)**	0.79 (0.79,0.80)	**0.76 (0.75,0.76)**	**0.87 (0.87,0.87)**	**0.83 (0.83,0.84)**	**0.88 (0.88,0.88)**
SOFA	0.73 (0.72,0.74)	0.60 (0.59,0.60)	0.75 (0.74,0.75)	0.70 (0.69,0.70)	0.73 (0.72,0.73)	0.70 (0.69,0.70)	0.69 (0.68,0.69)	**0.84 (0.83,0.84)**	0.67 (0.67,0.68)
SIRS	0.52 (0.51,0.53)	0.67 (0.66,0.69)	0.50 (0.49,0.52)	0.52 (0.51,0.53)	0.74 (0.73,0.75)	0.50 (0.49,0.51)	0.68 (0.68,0.69)	0.57 (0.56,0.57)	0.69 (0.69,0.70)
XG-SAPS	**0.77 (0.77,0.77)**	**0.75 (0.74,0.75)**	**0.78 (0.77,0.78)**	**0.81 (0.81,0.81)**	**0.84 (0.83,0.84)**	**0.81 (0.80,0.81)**	0.76 (0.75,0.77)	**0.92 (0.91,0.92)**	0.75 (0.74,0.76)
OASIS+	0.71 (0.71,0.72)	**0.78 (0.77,0.78)**	0.70 (0.70,0.71)	**0.80 (0.80,0.80)**	0.77 (0.76,0.78)	**0.80 (0.80,0.81)**	0.72 (0.71,0.72)	**0.83 (0.82,0.83)**	0.71 (0.70,0.71)
hXG-SAPS (w)	**0.77 (0.77,0.77)**	**0.75 (0.74,0.75)**	**0.77 (0.77,0.77)**	**0.81 (0.80,0.81)**	**0.86 (0.85,0.86)**	**0.80 (0.80,0.80)**	**0.78 (0.77,0.79)**	0.82 (0.81,0.83)	0.78 (0.77,0.79)
hXG-SAPS (u)	**0.77 (0.77,0.77)**	**0.75 (0.74,0.75)**	**0.77 (0.77,0.78)**	**0.80 (0.80,0.81)**	**0.85 (0.85,0.86)**	**0.80 (0.80,0.80)**	**0.81 (0.80,0.81)**	0.78 (0.77,0.79)	**0.81 (0.80,0.82)**
(S) hXG-SAPS (w)	0.74 (0.74,0.74)	0.69 (0.69,0.70)	0.75 (0.74,0.75)	0.71 (0.70,0.72)	0.73 (0.73,0.74)	0.71 (0.70,0.71)	0.74 (0.74,0.75)	**0.83 (0.82,0.84)**	0.74 (0.73,0.75)
(S) hXG-SAPS (u)	**0.77 (0.77,0.77)**	0.58 (0.58,0.58)	**0.79 (0.79,0.80)**	0.72 (0.72,0.73)	0.63 (0.63,0.64)	0.73 (0.73,0.73)	0.73 (0.72,0.73)	0.60 (0.59,0.61)	0.74 (0.73,0.74)

Method	Hispanic			Other		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
SAPS-II	0.67 (0.67,0.67)	0.86 (0.85,0.86)	**0.66 (0.65,0.66)**	0.70 (0.69,0.70)	0.74 (0.74,0.75)	0.69 (0.68,0.69)
OASIS	0.61 (0.60,0.61)	0.77 (0.76,0.78)	0.59 (0.59,0.60)	0.69 (0.69,0.70)	0.69 (0.68,0.70)	0.69 (0.68,0.70)
APS-III	**0.65 (0.64,0.65)**	0.68 (0.67,0.68)	0.64 (0.63,0.65)	**0.73 (0.73,0.73)**	0.71 (0.71,0.72)	**0.73 (0.73,0.74)**
SOFA	0.58 (0.57,0.59)	0.68 (0.67,0.69)	0.57 (0.56,0.58)	0.67 (0.66,0.67)	0.62 (0.62,0.63)	0.67 (0.67,0.68)
SIRS	0.54 (0.53,0.55)	0.60 (0.59,0.61)	0.54 (0.52,0.55)	0.64 (0.63,0.64)	0.51 (0.50,0.52)	0.66 (0.65,0.67)
XG-SAPS	0.63 (0.62,0.64)	**0.93 (0.93,0.94)**	0.61 (0.60,0.61)	**0.73 (0.73,0.73)**	**0.80 (0.80,0.81)**	**0.72 (0.71,0.72)**
OASIS+	**0.72 (0.71,0.72)**	0.85 (0.84,0.85)	**0.71 (0.70,0.71)**	**0.72 (0.71,0.72)**	0.74 (0.74,0.75)	**0.71 (0.71,0.71)**
hXG-SAPS (w)	0.63 (0.62,0.64)	**0.90 (0.89,0.91)**	0.61 (0.60,0.62)	**0.71 (0.70,0.71)**	**0.82 (0.81,0.82)**	0.69 (0.68,0.69)
hXG-SAPS (u)	**0.64 (0.63,0.65)**	**0.87 (0.87,0.88)**	0.62 (0.61,0.63)	**0.71 (0.71,0.72)**	**0.81 (0.80,0.81)**	0.70 (0.69,0.70)
(S) hXG-SAPS (w)	**0.72 (0.71,0.73)**	0.67 (0.66,0.68)	**0.73 (0.72,0.74)**	**0.71 (0.70,0.71)**	0.73 (0.72,0.73)	0.70 (0.70,0.71)
(S) hXG-SAPS (u)	**0.72 (0.71,0.73)**	0.56 (0.55,0.57)	**0.73 (0.72,0.74)**	**0.72 (0.72,0.73)**	0.53 (0.53,0.54)	0.76 (0.75,0.76)

Figure C.3: Performance summary of all predictive models applied to ethnicity stratified cohort. Accuracy, sensitivity, and specificity we report the mean of the bootstrapped evaluation and in parenthesis the 95% confidence interval. In bold we highlight the top three best-performing models for each evaluation metric

C.2 AUROC comparison of SAPS-II Machine Learning enhanced models: RF-SAPS and XG-SAPS

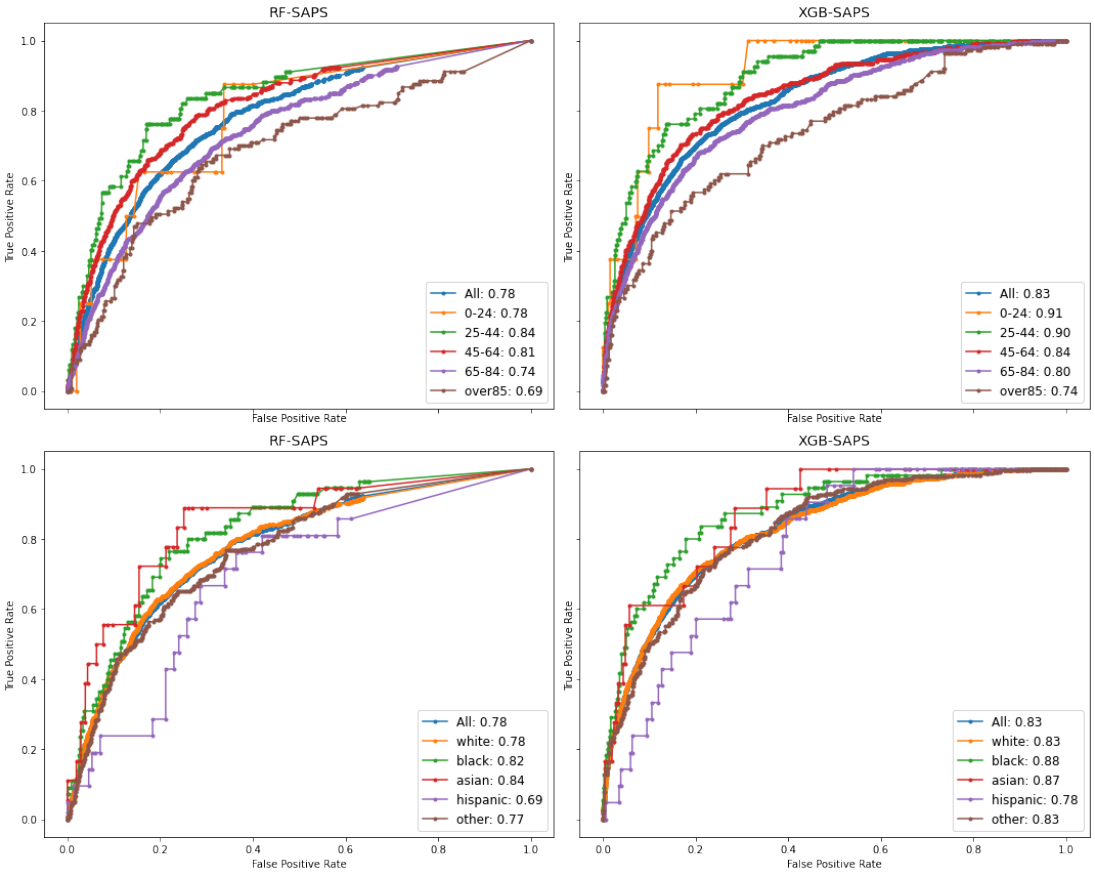

Figure C.4: AUROC comparison between RF-SAPS (left) and XG-SAPS (right). It can be seen that XG-SAPS outperforms RF-SAPS in all age (top) and ethnicity (bottom) cohorts. Based on these results we moved forward considering only XG-SAPS due to its superior performance.

Bibliography

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer.

Åkerlund, C. A., Holst, A., Stocchetti, N., Steyerberg, E. W., Menon, D. K., Ercole, A., and Nelson, D. W. (2022). Clustering identifies endotypes of traumatic brain injury in an intensive care cohort: a center-tbi study. *Critical Care*, 26(1):1–15.

Angus, D. C., Linde-Zwirble, W. T., Lidicker, J., Clermont, G., Carcillo, J., and Pinsky, M. R. (2001). Epidemiology of severe sepsis in the united states: analysis of incidence, outcome, and associated costs of care. *Critical care medicine*, 29(7):1303–1310.

Arnold, A. M., Psaty, B. M., Kuller, L. H., Burke, G. L., Manolio, T. A., Fried, L. P., Robbins, J. A., and Kronmal, R. A. (2005). Incidence of cardiovascular disease in older americans: The cardiovascular health study. *Journal of the American Geriatrics Society*, 53(2):211–218.

Austin, S. R., Wong, Y.-N., Uzzo, R. G., Beck, J. R., and Egleston, B. L. (2015). Why summary comorbidity measures such as the charlson comorbidity index and elixhauser score work. *Medical care*, 53(9):e65.

Ballard, H. S. (1997). The hematological complications of alcoholism. *Alcohol health and research world*, 21(1):42.

Barabási, A.-L. (2013). Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1987):20120375.

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature reviews genetics*, 12(1):56–68.

Barboi, C., Tzavelis, A., Muhammad, L. N., et al. (2022). Comparison of severity of illness scores and artificial intelligence models that are predictive of intensive care unit mortality: meta-analysis and review of the literature. *JMIR Medical Informatics*, 10(5):e35293.

Barnett, K., Mercer, S., Norbury, M., Watt, G., Wyke, S., and Guthrie, B. (2012a). Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *The Lancet*, 380:37–44.

Barnett, K., Mercer, S. W., Norbury, M., Watt, G., Wyke, S., and Guthrie, B. (2012b). Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *The Lancet*, 380(9836):37–43.

Barnett, K., Mercer, S. W., Norbury, M., Watt, G., Wyke, S., and Guthrie, B. (2012c). Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *The Lancet*, 380(9836):37–43.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, 33(2):261–304.

Busija, L., Lim, K., Szoeke, C., Sanders, K. M., and McCabe, M. P. (2019). Do replicable profiles of multimorbidity exist? systematic review and synthesis. *European journal of epidemiology*, 34(11):1025–1053.

Caliński, T. and Harabasz, J. (1974a). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.

Caliński, T. and Harabasz, J. (1974b). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.

Callahan, A. and Shah, N. H. (2017). Machine learning in healthcare. In *Key advances in clinical informatics*, pages 279–291. Elsevier.

Carioli, G., Bertuccio, P., Boffetta, P., Levi, F., La Vecchia, C., Negri, E., and Malvezzi, M. (2020). European cancer mortality predictions for the year 2020 with a focus on prostate cancer. *Annals of Oncology*, 31(5):650–658.

Charlson, M. E., Carrozzino, D., Guidi, J., and Patierno, C. (2022). Charlson comorbidity index: a critical review of clinimetric properties. *Psychotherapy and psychosomatics*, 91(1):8–35.

Charlson, M. E., Pompei, P., Ales, K. L., and MacKenzie, C. R. (1987). A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of chronic diseases*, 40(5):373–383.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Cohen, J., Vincent, J.-L., Adhikari, N. K., Machado, F. R., Angus, D. C., Calandra, T., Jaton, K., Giulieri, S., Delaloye, J., Opal, S., et al. (2015). Sepsis: a roadmap for future research. *The Lancet infectious diseases*, 15(5):581–614.

Cui, L., Xie, X., Shen, Z., Lu, R., and Wang, H. (2018). Prediction of the healthcare resource utilization using multi-output regression models. *IISE Transactions on Healthcare Systems Engineering*, 8(4):291–302.

Davies, M. and Hagen, P.-O. (1997). Systemic inflammatory response syndrome. *British Journal of Surgery*, 84(7):920–935.

Doreian, P., Batagelj, V., and Ferligoj, A. (2019). *Advances in Network Clustering and Blockmodeling*. John Wiley & Sons.

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pages 569–593. Springer.

El-Manzalawy, Y., Abbas, M., Hoaglund, I., Cerna, A. U., Morland, T. B., Haggerty, C. M., Hall, E. S., and Fornwalt, B. K. (2021). Oasis+: leveraging machine learning to improve the prognostic accuracy of oasis severity score for predicting in-hospital mortality. *BMC medical informatics and decision making*, 21(1):156.

El-Rashidy, N., El-Sappagh, S., Abuhmed, T., Abdelrazek, S., and El-Bakry, H. M. (2020). Intensive care unit mortality prediction: An improved patient-specific stacking ensemble model. *IEEE Access*, 8:133541–133564.

Elixhauser, A., Steiner, C., Harris, D. R., and Coffey, R. M. (1998). Comorbidity measures for use with administrative data. *Medical care*, pages 8–27.

Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., and Akinyelu, A. A. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, 110:104743.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.

Forte, J. C., Perner, A., and van der Horst, I. C. (2019). The use of clustering algorithms in critical care research to unravel patient heterogeneity. *Intensive care medicine*, pages 1–4.

Fortunato, S. and Hric, D. (2016). Community detection in networks: A user guide. *Physics reports*, 659:1–44.

Fuente-Tomas, L. d. l., Arranz, B., Safont, G., Sierra, P., Sanchez-Autet, M., Garcia-Blanco, A., and Garcia-Portilla, M. P. (2019). Classification of patients with bipolar disorder using k-means clustering. *PloS one*, 14(1):e0210314.

Funke, T. and Becker, T. (2019). Stochastic block models: A comparison of variants and inference methods. *PloS one*, 14(4):e0215296.

Geifman, N., Cohen, R., and Rubin, E. (2013). Redefining meaningful age groups in the context of disease. *Age*, 35(6):2357–2366.

Gerlach, M., Peixoto, T. P., and Altmann, E. G. (2018). A network approach to topic models. *Science advances*, 4(7):eaaq1360.

Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., Chen, I. Y., and Ranganath, R. (2020). A review of challenges and opportunities in machine learning for health. *AMIA Summits on Translational Science Proceedings*, 2020:191.

Grant, R. W., McCloskey, J., Hatfield, M., Uratsu, C., Ralston, J. D., Bayliss, E., and Kennedy, C. J. (2020). Use of latent class analysis and k-means clustering to identify complex patient profiles. *JAMA network open*, 3(12):e2029068–e2029068.

Hagenaars, J. A. and McCutcheon, A. L. (2002). *Applied latent class analysis*. Cambridge University Press.

Hartmann, H. C., Pagano, T. C., Sorooshian, S., and Bales, R. (2002). Confidence builders: Evaluating seasonal climate forecasts from user perspectives. *Bulletin of the American Meteorological Society*, 83(5):683–698.

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137.

Hu, J., Perer, A., and Wang, F. (2016). Data driven analytics for personalized healthcare. *Healthcare Information Management Systems: Cases, Strategies, and Solutions*, pages 529–554.

Huang, Z. (1997). A fast clustering algorithm to cluster very large categorical data sets in data mining. *Dmkd*, 3(8):34–39.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, 2(3):283–304.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., Lauer, M. S., et al. (2008). Random survival forests. *The annals of applied statistics*, 2(3):841–860.

Iwashyna, T. J., Odden, A., Rohde, J., Bonham, C., Kuhn, L., Malani, P., Chen, L., and Flanders, S. (2014). Identifying patients with severe sepsis using administrative claims: patient-level validation of the angus implementation of the international consensus conference definition of severe sepsis. *Medical care*, 52(6):e39.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323.

Javed, M. A., Younis, M. S., Latif, S., Qadir, J., and Baig, A. (2018). Community detection in networks: A multidisciplinary review. *Journal of Network and Computer Applications*, 108:87–111.

Johnson, A., Pollard, T., and Mark, R. (2016a). Mimic-iii clinical database (version 1.4). *PhysioNet*, 10(C2XW26):2.

Johnson, A. E., Ghassemi, M. M., Nemati, S., Niehaus, K. E., Clifton, D. A., and Clifford, G. D. (2016b). Machine learning and decision support in critical care. *Proceedings of the IEEE*, 104(2):444–466.

Johnson, A. E., Kramer, A. A., and Clifford, G. D. (2013a). A new severity of illness scale using a subset of acute physiology and chronic health evaluation data elements shows comparable predictive accuracy. *Critical care medicine*, 41(7):1711–1718.

Johnson, A. E., Kramer, A. A., and Clifford, G. D. (2013b). A new severity of illness scale using a subset of acute physiology and chronic health evaluation data elements shows comparable predictive accuracy. *Critical care medicine*, 41(7):1711–1718.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. (2016c). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Johnson, A. E., Pollard, T. J., Shen, L., Li-Wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016d). MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Johnson, A. E. W., Stone, D. J., Celi, L. A., and Pollard, T. J. (2018). The MIMIC code repository: enabling reproducibility in critical care research. *Journal of the American Medical Informatics Association*, 25(1):32–39.

Kalgotra, P., Sharda, R., and Croff, J. M. (2017). Examining health disparities by gender: A multimorbidity network analysis of electronic medical record. *International journal of medical informatics*, 108:22–28.

Kalgotra, P., Sharda, R., and Croff, J. M. (2020). Examining multimorbidity differences across racial groups: a network analysis of electronic medical records. *Scientific reports*, 10(1):1–9.

Karrer, B. and Newman, M. E. (2011). Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1):016107.

Keuning, B. E., Kaufmann, T., Wiersema, R., Granholm, A., Pettilä, V., Møller, M. H., Christiansen, C. F., Castela Forte, J., Snieder, H., Keus, F., et al. (2020). Mortality prediction models in the adult critically ill: A scoping review. *Acta Anaesthesiologica Scandinavica*, 64(4):424–442.

Kong, G., Lin, K., and Hu, Y. (2020). Using machine learning methods to predict in-hospital mortality of sepsis patients in the icu. *BMC medical informatics and decision making*, 20:1–10.

Le Gall, J.-R., Lemeshow, S., and Saulnier, F. (1993). A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *Jama*, 270(24):2957–2963.

Lin, K., Hu, Y., and Kong, G. (2019). Predicting in-hospital mortality of patients with acute kidney injury in the icu using random forest model. *International journal of medical informatics*, 125:55–61.

Linzer, D. A. and Lewis, J. B. (2011). polca: An r package for polytomous variable latent class analysis. *Journal of statistical software*, 42:1–29.

Lvovschi, V., Arnaud, L., Parizot, C., Freund, Y., Juillien, G., Ghillani-Dalbin, P., Bouberima, M., Larsen, M., Riou, B., Gorochov, G., et al. (2011). Cytokine profiles in sepsis have limited relevance for stratifying patients in the emergency department: a prospective observational study. *PloS one*, 6(12):e28870.

Martin, T., Ball, B., and Newman, M. E. (2016). Structural inference for uncertain networks. *Physical Review E*, 93(1):012306.

Maslove, D. M., Lamontagne, F., Marshall, J. C., and Heyland, D. K. (2017). A path to precision in the icu. *Critical Care*, 21(1):1–9.

Metnitz, P. G., Moreno, R. P., Almeida, E., Jordan, B., Bauer, P., Campos, R. A., Iapichino, G., Edbrooke, D., Capuzzo, M., Le Gall, J.-R., et al. (2005). Saps 3—from evaluation of the patient to evaluation of the intensive care unit. part 1: Objectives, methods and cohort description. *Intensive care medicine*, 31:1336–1344.

Michiels, B., Molenberghs, G., Bijnens, L., Vangeneugden, T., and Thijs, H. (2002). Selection models and pattern-mixture models to analyse longitudinal quality of life data subject to drop-out. *Statistics in medicine*, 21(8):1023–1041.

Moreno, R. and Apolone, G. (1997). Impact of different customization strategies in the performance of a general severity score. *Critical care medicine*, 25(12):2001–2008.

Moreno, R. P., Metnitz, P. G., Almeida, E., Jordan, B., Bauer, P., Campos, R. A., Iapichino, G., Edbrooke, D., Capuzzo, M., Le Gall, J.-R., et al. (2005). Saps 3—from evaluation of the patient to evaluation of the intensive care unit. part 2: Development of a prognostic model for hospital mortality at icu admission. *Intensive care medicine*, 31:1345–1355.

Moshkovitz, M., Dasgupta, S., Rashtchian, C., and Frost, N. (2020). Explainable k-means and k-medians clustering. In *International Conference on Machine Learning*, pages 7055–7065. PMLR.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Nasserinejad, K., van Rosmalen, J., de Kort, W., and Lesaffre, E. (2017). Comparison of criteria for choosing the number of classes in bayesian finite mixture models. *PloS one*, 12(1):e0168838.

NCHS (1980). *The International classification of diseases, 9th revision, clinical modification: ICD-9-CM*, volume 2. US Department of Health and Human Services, Public Health Service, Health ....

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582.

Newman, M. E. J. (2018). *Networks*. Oxford university press.

Papachristou, N., Barnaghi, P., Cooper, B. A., Hu, X., Maguire, R., Apostolidis, K., Armes, J., Conley, Y. P., Hammer, M., Katsaragakis, S., et al. (2018). Congruence between latent class and k-modes analyses in the identification of oncology patients with distinct symptom experiences. *Journal of pain and symptom management*, 55(2):318–333.

Papin, G., Bailly, S., Dupuis, C., Ruckly, S., Gainnier, M., Argaud, L., Azoulay, E., Adrie, C., Souweine, B., Goldgran-Toledano, D., et al. (2021). Clinical and biological clusters of sepsis patients using hierarchical clustering. *PloS one*, 16(8):e0252793.

Pearce, C. B., Gunn, S. R., Ahmed, A., and Johnson, C. D. (2006). Machine learning can improve prediction of severity in acute pancreatitis using admission values of apache ii score and c-reactive protein. *Pancreatology*, 6(1-2):123–131.

Peixoto, T. P. (2014a). Efficient monte carlo and greedy heuristic for the inference of stochastic block models. *Physical Review E*, 89:012804.

Peixoto, T. P. (2014b). Hierarchical block structures and high-resolution model selection in large networks. *Physical Review X*, 4(1):011047.

Peixoto, T. P. (2017). Nonparametric bayesian inference of the microcanonical stochastic block model. *Physical Review E*, 95(1):012317.

Peixoto, T. P. (2019). Bayesian stochastic blockmodeling. *Advances in network clustering and blockmodeling*, pages 289–332.

Peixoto, T. P. (2020). Merge-split markov chain monte carlo for community detection. *Physical Review E*, 102(1):012305.

Peixoto, T. P. (2021). Revealing consensus and dissensus between network partitions. *Physical Review X*, 11(2):021003.

Popoola, P. A., Tapamo, J.-R., and Assounga, A. G. (2021). Cluster analysis of mixed and missing chronic kidney disease data in kwazulu-natal province, south africa. *IEEE Access*, 9:52125–52143.

Prados-Torres, A., Calderón-Larrañaga, A., Hancco-Saavedra, J., Poblador-Plou, B., and van den Akker, M. (2014). Multimorbidity patterns: a systematic review. *Journal of clinical epidemiology*, 67(3):254–266.

Raykov, Y. P., Boukouvalas, A., Baig, F., and Little, M. A. (2016). What to do when k-means clustering fails: a simple yet principled alternative algorithm. *PloS one*, 11(9):e0162259.

Reddy, K., Sinha, P., O'Kane, C. M., Gordon, A. C., Calfee, C. S., and McAuley, D. F. (2020). Subphenotypes in critical care: translation into clinical practice. *The Lancet Respiratory Medicine*, 8(6):631–643.

Restocchi, V., Villegas, J. G., and Fleuriot, J. D. (2022). Multimorbidity profiles and stochastic block modeling improve icu patient clustering. In *2022 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*, pages 925–932. IEEE.

Rindskopf, D. and Rindskopf, W. (1986). The value of latent class analysis in medical diagnosis. *Statistics in medicine*, 5(1):21–27.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14(5):465–471.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Rumelhart, D. E., Hinton, G. E., Williams, R. J., et al. (1985). Learning internal representations by error propagation.

Salmon-Ceron, D., Lewden, C., Morlat, P., Bévilacqua, S., Jougla, E., Bonnet, F., Héripret, L., Costagliola, D., May, T., and Chêne, G. (2005). Liver disease as a major cause of death among HIV infected patients: role of hepatitis C and B viruses and alcohol. *Journal of Hepatology*, 42(6):799–805.

Satopaa, V., Albrecht, J., Irwin, D., and Raghavan, B. (2011). Finding a" kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st international conference on distributed computing systems workshops*, pages 166–171. IEEE.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, pages 461–464.

Seymour, C. W., Kennedy, J. N., Wang, S., Chang, C.-C. H., Elliott, C. F., Xu, Z., Berry, S., Clermont, G., Cooper, G., Gomez, H., et al. (2019). Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. *JAMA*, 321(20):2003–2017.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., and Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, 338.

Strand, K., Søreide, E., Aardal, S., and Flaatten, H. (2009). A comparison of saps ii and saps 3 in a norwegian intensive care unit population. *Acta anaesthesiologica scandinavica*, 53(5):595–600.

Teh, R. O., Menzies, O. H., Connolly, M. J., Doughty, R. N., Wilkinson, T. J., Pillai, A., Lumley, T., Ryan, C., Rolleston, A., Broad, J. B., et al. (2018). Patterns of multimorbidity and prediction of hospitalisation and all-cause mortality in advanced age. *Age and ageing*, 47(2):261–268.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.

Trevithick, L., Painter, J., and Keown, P. (2015). Mental health clustering and diagnosis in psychiatric in-patients. *BJPsych Bulletin*, 39(3):119–123.

Vellido, A. (2020a). The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural computing and applications*, 32(24):18069–18083.

Vellido, A. (2020b). The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural computing and applications*, 32(24):18069–18083.

Venugopalan, J., Chanani, N., Maher, K., and Wang, M. D. (2019). Novel data imputation for multiple types of missing data in intensive care units. *IEEE journal of biomedical and health informatics*, 23(3):1243–1250.

Vesin, A., Azoulay, E., Ruckly, S., Vignoud, L., Rusinovà, K., Benoit, D., Soares, M., Azeivedo-Maia, P., Abroug, F., Benbenishty, J., et al. (2013). Reporting and handling missing values in clinical studies in intensive care units. *Intensive care medicine*, 39:1396–1404.

Vincent, J.-L. and Moreno, R. (2010). Clinical review: scoring systems in the critically ill. *Critical care*, 14:1–9.

Vincent, J.-L., Moreno, R., Takala, J., Willatts, S., De Mendonça, A., Bruining, H., Reinhart, C., Suter, P., and Thijs, L. G. (1996a). The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure.

Vincent, J. L., Moreno, R., Takala, J., Willatts, S., De Mendonça, A., Bruining, H., Reinhart, C., Suter, P., and Thijs, L. G. (1996b). The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure: On behalf of the working group on sepsis-related problems of the european society of intensive care medicine (see contributors to the project in the appendix).

Walsh, C. G., Sharman, K., and Hripcsak, G. (2017). Beyond discrimination: a comparison of calibration methods and clinical usefulness of predictive models of readmission risk. *Journal of biomedical informatics*, 76:9–18.

Wilson, P. W., Kannel, W. B., Silbershatz, H., and D'Agostino, R. B. (1999). Clustering of metabolic factors and coronary heart disease. *Archives of internal medicine*, 159(10):1104–1109.

Xu, D. and Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2:165–193.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1):32–35.

Zador, Z., Landry, A., Cusimano, M. D., and Geifman, N. (2019). Multimorbidity states associated with higher mortality rates in organ dysfunction and sepsis: a data-driven analysis in critical care. *Critical Care*, 23(1):1–11.

Zhang, Z., Zhang, G., Goyal, H., Mo, L., and Hong, Y. (2018). Identification of subclasses of sepsis that showed different clinical outcomes and responses to amount of fluid resuscitation: a latent profile analysis. *Critical Care*, 22(1):1–11.

Zimmerman, J. E., Kramer, A. A., McNair, D. S., and Malila, F. M. (2006). Acute physiology and chronic health evaluation (apache) iv: hospital mortality assessment for today's critically ill patients. *Critical care medicine*, 34(5):1297–1310.