

Phenotyping Obstructive Sleep Apnea: integration of clinical and biomedical data through clustering analysis

Lara Filipa Faria Miranda

Tese de Mestrado

Supervisor: Prof. Doutora Daniela Ferreira-Santos



Mestrado em Engenharia Biomédica

2026-06-21

Resumo

A Apneia Obstrutiva do Sono (AOS) é uma condição crónica de respiração desordenada durante o sono, caracterizada por episódios recorrentes de colapso parcial ou total das vias aéreas superiores, afetando aproximadamente 936 milhões de adultos a nível global. O seu impacto clínico é multissistémico, abrangendo consequências cardiovasculares, metabólicas e neurocognitivas que transcendem a métrica clássica de severidade baseada no Índice de Apneia-Hipopneia (IAH). O reconhecimento crescente de que doentes com valores idênticos de IAH podem apresentar perfis clínicos e respostas terapêuticas fundamentalmente distintas motivou a procura por estratégias de estratificação fenotípica baseadas em dados clínicos de pré-triagem. Esta dissertação teve como objetivo principal identificar e caracterizar fenótipos clínicos de AOS utilizando o algoritmo k-modes, aplicado exclusivamente a variáveis categóricas de rotina clínica, visando promover uma medicina de precisão e otimizar a triagem.

Foram analisadas duas coortes independentes provenientes de centros terciários ibéricos: uma coorte portuguesa ($n = 188$; 78 saudáveis, 110 AOS) e uma coorte espanhola ($n = 355$; 68 saudáveis, 287 AOS), abrangendo 23 variáveis clínicas categóricas. O k-modes foi aplicado sob duas estratégias complementares de seleção de variáveis: uma abordagem clínica, baseada em seleção univariada guiada pelo diagnóstico (retendo as variáveis com $p < 0.05$), e uma abordagem científica, puramente orientada pelos dados, incorporando a totalidade das variáveis sem supervisão diagnóstica. A abordagem clínica identificou seis clusters na coorte portuguesa, organizados ao longo dos eixos de índice de massa corporal (IMC), hipertensão e insónia, com valores médios de IAH variando entre 14.9 ± 16.9 (fenótipo de baixa severidade) e 33.3 ± 22.3 (fenótipo de alta severidade). Na coorte espanhola, quatro clusters emergiram organizados por idade, hipertensão e diabetes, com IAH médio entre 34.3 ± 32.7 e 56.6 ± 32.8 . A abordagem científica convergiu para uma solução de dois clusters em ambas as coortes, revelando um padrão fundamental de variabilidade clínica associado à relação comorbilidade idade, conservado entre populações. Na coorte espanhola, os dois clusters apresentaram distribuições de IAH virtualmente idênticas (média 44.4 ± 31.3 vs. 44.8 ± 28.5), apesar de perfis clínicos diametralmente opostos: um cluster sintomático com baixa carga cardiometabólica e um cluster assintomático com elevada carga cardiovascular.

Esta dissociação demonstra empiricamente que a classificação baseada exclusivamente no IAH negligencia eixos cruciais da variância fenotípica, nomeadamente a carga cardiometabólica, a idade e a apresentação sintomática do doente. Adicionalmente, implementaram-se gráficos de radar juntamente com mapas de calor, facilitando o reconhecimento visual rápido destes perfis específicos, como as assinaturas geométricas contrastantes entre fenótipos altamente comórbidos e fenótipos puramente sintomáticos, que as representações tabulares tradicionais frequentemente ocultam.

Estes resultados confirmam a viabilidade e a utilidade clínica da fenotipagem pré-diagnóstico, demonstrando que a caracterização multidimensional com dados categóricos de rotina tem o potencial de refinar os percursos de triagem diagnóstica e informar intervenções terapêuticas mais personalizadas na abordagem à AOS.

Palavras-chave: Apneia Obstrutiva do Sono, Fenotipagem Clínica, Clustering de Dados Categóricos, k-modes, Medicina de Precisão, Visualização de Dados Médicos.

Abstract

Obstructive Sleep Apnea (OSA) is a chronic sleep-disordered breathing condition characterized by recurrent episodes of partial or complete upper airway collapse during sleep, affecting an estimated 936 million adults worldwide. Its clinical impact is multisystemic, encompassing cardiovascular, metabolic, and neurocognitive consequences that extend well beyond the classical severity metric based on the Apnea-Hypopnea Index (AHI). The growing recognition that patients with identical AHI values can present fundamentally different clinical profiles and therapeutic responses has driven the search for phenotypic stratification strategies grounded in pre-triage clinical data. The primary objective of this dissertation was to identify and characterize clinical OSA phenotypes using the k-modes clustering algorithm applied exclusively to routine categorical clinical variables, aiming to advance precision medicine and optimize triage.

Two independent cohorts from Iberian tertiary centers were analyzed: a Portuguese cohort ($n = 188$; 78 healthy, 110 OSA) and a Spanish cohort ($n = 355$; 68 healthy, 287 OSA), encompassing 23 categorical clinical variables. The k-modes algorithm was applied under two complementary variable selection strategies: a clinical approach, based on univariate selection guided by the diagnostic label (retaining variables with $p < 0.05$), and a scientific approach, a purely data-driven method incorporating all available variables without diagnostic supervision. The clinical approach identified six clusters in the Portuguese cohort, organized along the body mass index (BMI), hypertension, and insomnia axes, with mean AHI values ranging from 14.9 ± 16.9 (low-severity phenotype) to 33.3 ± 22.3 (high-severity phenotype). In the Spanish cohort, four clusters emerged organized by age, hypertension, and diabetes, with mean AHI ranging from 34.3 ± 32.7 to 56.6 ± 32.8 . The scientific approach converged on a two-cluster solution in both cohorts, revealing a fundamental pattern of clinical variability associated with the comorbidity age relationship, conserved across populations. In the Spanish cohort, the two clusters exhibited virtually identical AHI distributions (mean 44.4 ± 31.3 vs. 44.8 ± 28.5) despite diametrically opposed clinical profiles: one symptomatic cluster with low cardiometabolic burden and one asymptomatic cluster with extensive cardiovascular burden.

This dissociation empirically demonstrates that AHI-based classification alone neglects critical axes of phenotypic variance, specifically the patient's cardiometabolic burden, age, and symptomatic presentation. Additionally, radar plots were employed alongside traditional heatmaps, facilitating the rapid visual recognition of these specific profiles, such as the contrasting geometric signatures of highly comorbid versus highly symptomatic phenotypes which traditional tabular data often obscure.

These findings confirm the feasibility and clinical utility of pre-diagnostic phenotyping, demonstrating that multidimensional characterization using routine categorical data has the potential to refine diagnostic triage pathways and inform personalized therapeutic interventions in OSA management.

Keywords: Obstructive Sleep Apnea, Clinical Phenotyping, Categorical Data Clustering, k-modes, Precision Medicine, Medical Data Visualization.

The United Nations Sustainable Development Goals

The research developed in this dissertation is fundamentally aligned with the United Nations Agenda 2030 ¹, contributing directly to the Sustainable Development Goals (SDGs). By addressing the diagnostic bottlenecks and phenotypic heterogeneity of Obstructive Sleep Apnea (OSA) through a data-driven precision medicine framework, this project generates impact across three primary SDGs: Good Health and Well-being (SDG 3), Gender Equality (SDG 5), and Reduced Inequalities (SDG 10).

The following table details the specific targets addressed by this work. The transition from a one-dimensional diagnostic paradigm to a phenotype-driven framework not only improves individual patient care but also addresses broader societal challenges, including gender bias in medical research and disparities in healthcare access.

SDG	Target	Contribution	Performance indicators and metrics
3	3.4	Good Health and Well-being: By improving the phenotyping of OSA patients using readily available clinical data, this research contributes to earlier triage and personalized treatment, ultimately helping reduce cardiovascular and metabolic morbidity associated with untreated OSA.	Identification of specific high-risk cardiometabolic phenotypes; development of clinically actionable cluster profiles.
5	5.1	Gender Equality: The study addresses the historical male-biased paradigm in OSA by explicitly identifying sex-specific phenotypic expressions. Demonstrating that women present distinct clinical profiles (e.g., an older, highly comorbid phenotype) combats historical underdiagnosis and promotes gender equity in sleep medicine.	Analysis of sex distribution across identified clusters; characterization of the distinct female clinical profile versus the classical male presentation.
10	10.2	Reduced Inequalities: The reliance on expensive in-lab polysomnography creates diagnostic bottlenecks that disproportionately affect lower-income populations. Showing that meaningful clinical stratification can be achieved using basic pre-triage variables democratizes triage, promoting more equitable healthcare access.	Demonstration of the discriminative power of basic pre-triage clinical data without relying on costly polysomnographic features.

¹United Nations. *Transforming our world: the 2030 Agenda for Sustainable Development*. General Assembly Resolution 70/1 (2015). Available at: <https://sdgs.un.org/2030agenda>

Acknowledgments

First, I would like to thank my supervisor Daniela Ferreira-Santos. Without her support, this work would not have been possible. I appreciate her availability and the knowledge she shared with me. I greatly admire her good spirit, which made the challenges of this journey much easier to overcome.

I would also like to thank our Spanish partners and the team that contributed to this work, especially Javier Gomez-Pilar. They were always tireless and willing to answer all my questions.

To my family, my sincere thanks for always being present and for believing in me. Especially to my parents, Beatriz and António: I am extremely grateful that you provided me with this opportunity. I will carry your values with me forever, and I hope you are as proud of me as I am proud to be your daughter.

To my sister, Licínia, who is always there when I need her. You are my greatest inspiration and a woman I admire very much. My whole life, I have been a mosaic of you.

To Jon for his support and to my godchildren, who inspire me every day to become a better person for their future.

I also want to thank my friends, who were present throughout these years, you became home to me and made this journey more special, filled with memories that will stay with me forever.

To Gonçalo, thank you for your love, your patience, for always being by my side, and for encouraging me to do more and be better! Thank you for being my safe haven and for bringing me peace during the most stressful days.

Finally, to everyone who contributed directly or indirectly to this work, a massive thank you.

Declaration of Honor

I hereby declare that this dissertation is my own work and has not previously been submitted for assessment at this or any other institution. References to other authors, including statements, ideas and thoughts, as well as the use of published works, strictly comply with the rules of attribution and are duly indicated in the text and in the bibliographic references, in accordance with the applicable referencing standards. I am aware that the practice of plagiarism or self-plagiarism constitutes academic misconduct, as established in the Code of Ethics of the University of Porto.

I further declare that Artificial Intelligence tools were used as auxiliary support in the preparation of this dissertation, namely for tasks such as language refinement, text structuring and support in the revision of written content. Their use did not replace the authorial, scientific or analytical work developed in this dissertation. All outputs obtained through these tools were carefully reviewed and critically assessed in order to detect and correct possible errors, inaccuracies, unfounded attributions, biased content or arguments, and to ensure the appropriate determination of authorship.

Lara Filipa Faria Miranda



Contents

1	Introduction	1
1.1	Context and motivation	1
1.2	Obstructive sleep apnea: definition and clinical impact	2
1.3	Phenotypic heterogeneity and the need for stratification	3
1.4	Dissertation structure	5
2	Literature review	6
2.1	Introduction	6
2.2	Clustering methods in OSA research	6
2.2.1	Overview of traditional algorithms	6
2.2.2	K-modes clustering for categorical clinical data	7
2.2.3	Phenotypic findings from traditional clustering	8
2.2.4	Comparative algorithm evaluation	10
2.3	Emerging methods: network based and multimodal clustering	10
2.4	Limitations in current research	12
2.5	Objectives	12
3	Methods	14
3.1	Study design and population	14
3.2	Variables and data description	14
3.2.1	Variable categorization	15
3.2.2	Variable inventory	15
3.3	Analytical approaches	16
3.3.1	Approach 1: clinically guided feature selection	16
3.3.2	Approach 2: data driven full variable inclusion	17
3.3.3	Software and implementation	17
4	Results	19
4.1	Clinical approach	19
4.1.1	Portuguese dataset	19
4.1.2	Spanish dataset	24
4.2	Scientific approach	30
4.2.1	Portuguese dataset	30
4.2.2	Spanish dataset	34
4.2.3	Comparative radar plot analysis	38
5	Discussion	39
5.1	Cohort comparison and selection bias	39
5.1.1	Inter-cohort discrepancies	39
5.1.2	Referral bias and diagnostic pathway asymmetry	40
5.2	Phenotypic interpretation and comparison with the literature	41
5.2.1	Clinical approach: cohort-specific phenotypic drivers	41
5.2.2	Scientific approach: cross-cohort convergence on a comorbidity–age axis	42
5.2.3	Cross-cohort phenotypic archetypes	44
5.3	Confrontation of clustering approaches	46
5.4	Limitations	47

6 Conclusion	49
6.1 Future directions	50
References	51

Acronyms

AASM	American Academy of Sleep Medicine
AHI	Apnea-Hypopnea Index
AUC	Area Under the Curve
BMI	Body Mass Index
CBT-I	Cognitive Behavioral Therapy for Insomnia
CHI	Calinski-Harabasz Index
COMISA	Co-morbid Insomnia and Sleep Apnea
COPD	Chronic Obstructive Pulmonary Disease
CPAP	Continuous Positive Airway Pressure
DBI	Davies-Bouldin Index
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DI	Dunn Index
EDS	Excessive Daytime Sleepiness
EEG	Electroencephalography
ESADA	European Sleep Apnea Database
ESS	Epworth Sleepiness Scale
GE	Gastroesophageal
HRV	Heart Rate Variability
HSAT	Home Sleep Apnea Testing
LCA	Latent Class Analysis
OSA	Obstructive Sleep Apnea
PAP	Positive Airway Pressure
PSG	Polysomnography
QUADAS	Quality Assessment of Diagnostic Accuracy Studies
REM	Rapid Eye Movement
SBII	Sleep Breathing Impairment Index
UW	Underweight
WHO	World Health Organization

List of Figures

1	Schematic comparison of a normal and an obstructed upper airway during an OSA event. Adapted from Tan et al. (2021) [9].	2
2	Conceptual framework for the transition from AHI-centric diagnosis to phenotype-driven precision medicine (Author's own elaboration).	5
3	Heatmap of categorical variable distributions across the six clusters identified in the Portuguese cohort (clinical approach).	22
4	Radar plot of the six clusters identified in the Portuguese cohort (clinical approach). Each polygon represents a cluster's multivariate profile across the discriminating variables.	23
5	Heatmap of categorical variable distributions across the four clusters identified in the Spanish cohort (clinical approach).	28
6	Radar plot of the four clusters identified in the Spanish cohort (clinical approach). Each polygon represents a cluster's multivariate profile across the discriminating variables.	29
7	Heatmap of categorical variable distributions across the two clusters identified in the Portuguese cohort (scientific approach).	33
8	Heatmap of categorical variable distributions across the two clusters identified in the Spanish cohort (scientific approach).	37
9	Radar plots generated for both the Portuguese (left) and Spanish (right) two-cluster solutions. Each polygon represents a cluster's multivariate profile across all 23 clinical variables.	38

List of Tables

1	Descriptive characteristics of the Portuguese study cohort, n (%). P-values obtained using the chi square test unless otherwise specified.	20
2	Clinical characteristics of the Portuguese cohort by the defined clusters, n (%). . .	21
3	Modes of the clustered clinical presentations of OSA – Portuguese cohort (clinical approach).	22
4	Descriptive characteristics of the Spanish study cohort, n (%). P-values obtained using the chi square test unless otherwise specified.	25
5	Clinical characteristics of the Spanish cohort by the defined clusters, n (%). . . .	26
6	Modes of the clustered clinical presentations of OSA – Spanish cohort (clinical approach).	27
7	Clinical characteristics of the Portuguese cohort by the defined clusters (scientific approach), n (%).	31
8	Modes of the clustered clinical presentations of OSA – Portuguese cohort (scientific approach).	32
9	Clinical characteristics of the Spanish cohort by the defined clusters (scientific approach), n (%).	35
10	Modes of the clustered clinical presentations of OSA – Spanish cohort (scientific approach).	36

Chapter 1

Introduction

1.1 Context and motivation

Obstructive Sleep Apnea (OSA) is a chronic sleep disordered breathing condition that constitutes one of the most prevalent yet underdiagnosed disorders in clinical sleep medicine [1; 2]. Global epidemiological estimates indicate that approximately 936 million adults aged 30 to 69 years are affected by mild to severe forms of the condition, of whom 425 million present moderate to severe disease [1]. From a sex stratified perspective, OSA affects approximately 4% of adult men and at least 2% of adult women, although more recent population based studies suggest these figures are likely conservative due to historically male biased sampling [1; 3]. The disorder is not merely a matter of disturbed sleep: untreated OSA is independently associated with a significantly elevated risk of cardiovascular events, metabolic dysfunction, neurocognitive impairment, and all cause mortality [4].

From a socioeconomic perspective, the burden of OSA extends far beyond direct healthcare costs. Excessive daytime sleepiness, a cardinal symptom of the disorder, is causally linked to a heightened risk of motor vehicle and occupational accidents, with systematic reviews and meta analyses demonstrating a statistically significant odds ratio for traffic accidents in affected populations [2; 5]. Furthermore, the indirect costs attributable to reduced workplace productivity, absenteeism, and presenteeism constitute a substantial economic drain that disproportionately affects working age adults [2]. Despite these consequences, the diagnostic pathway remains constrained by the reliance on in-laboratory polysomnography (in-lab PSG), which is the gold standard for OSA diagnosis [6]. In-lab PSG provides comprehensive and reliable physiological measurements, including electroencephalography (EEG), airflow signals, electromyography, and peripheral oxygen saturation (SpO_2), but it is resource intensive, requiring specialized facilities, trained technicians, and an overnight patient stay. This resource burden creates significant diagnostic bottlenecks, resulting in long waiting lists and leaving a large proportion of the affected population undiagnosed [6; 7].

The confluence of high prevalence, severe clinical consequences, and diagnostic limitations has spurred a growing interest in alternative strategies to triage and stratify patients at risk for OSA before referral for full overnight sleep studies. The motivation behind this project is grounded in a precision medicine framework, aiming to improve OSA care by identifying clinically meaningful patient clusters through data driven integration of clinical and biomedical information. Specifically, if distinct phenotypic profiles can be reliably identified using pre-triage clinical variables, including demographics, self reported symptoms, anthropometric measures, and comorbidity patterns, then it becomes possible to prioritize patients for diagnostic testing, personalize therapeutic

strategies, and ultimately reduce the socioeconomic burden of the disorder [4; 6]. Beyond individual clinical improvements, the core motivation behind this dissertation is driven by a broader societal imperative: actively contributing to the United Nations Sustainable Development Goals (SDGs) [8].

1.2 Obstructive sleep apnea: definition and clinical impact

OSA is characterized by recurrent episodes of partial (hypopnea) or complete (apnea) upper airway collapse during sleep, as schematically illustrated in Fig. 1. These obstructive events lead to intermittent hypoxemia, that is, cyclical drops in arterial oxygen saturation, and sleep fragmentation due to cortical arousals triggered by the body's effort to restore airway patency [1; 4]. The repetitive cycles of desaturation and reoxygenation induce oxidative stress and sympathetic activation, which are recognized as central pathophysiological mediators linking OSA to its systemic consequences [4].

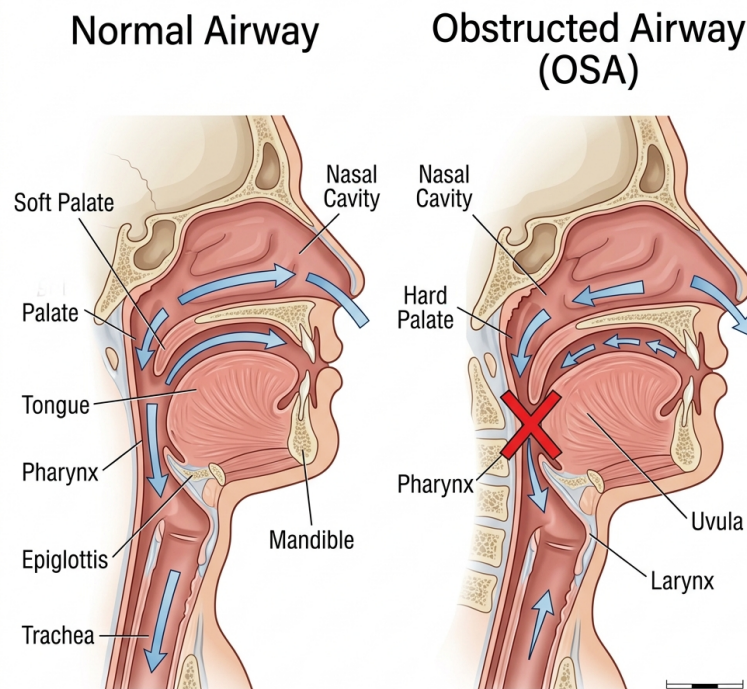


Figure 1: Schematic comparison of a normal and an obstructed upper airway during an OSA event. Adapted from Tan et al. (2021) [9].

The clinical impact of OSA is multisystemic. The cardiovascular consequences are among the most extensively documented, with strong evidence linking OSA to arterial hypertension, coronary artery disease, cardiac arrhythmias, heart failure, and stroke [10; 11]. Zinchuk et al. (2018) [11] demonstrated that polysomnographic phenotypes of OSA carry differential cardiovascular implications, with certain subtypes exhibiting significantly higher odds of incident cardiovascular

events even after adjusting for AHI. Metabolic dysfunction is another critical consequence: OSA is independently associated with insulin resistance, type 2 diabetes, and dyslipidemia, with the severity of intermittent hypoxemia serving as a stronger metabolic predictor than the AHI alone [12]. Neurocognitive sequelae include impaired vigilance, executive dysfunction, memory deficits, and mood disturbances, all of which contribute to a diminished quality of life and functional capacity [2]. Yook et al. (2025) [13] further showed that OSA subtypes based on hypoxic burden are differentially associated with brain aging markers and cardiometabolic syndrome, underscoring that the downstream effects of OSA are not monolithic but phenotype dependent.

The diagnosis is established and quantified using the Apnea-Hypopnea Index (AHI), defined as the count of the number of respiratory events (apneas and hypopneas) per hour of sleep, measured via in-lab PSG or, in selected cases, home based polysomnography [14; 15]. A diagnosis is established when a patient has an AHI of 5 or greater with associated symptoms, or an AHI of 15 or greater regardless of associated symptoms [1]. Standard diagnostic criteria further classify OSA severity into mild (≥ 5 events/hour), moderate (≥ 15 events/hour), and severe (≥ 30 events/hour). However, the reliance on the AHI as the sole arbiter of disease severity constitutes a well recognized and increasingly criticized limitation of the current diagnostic paradigm [4].

The fundamental inadequacy of the AHI resides in its reductive nature: it collapses a complex, multidimensional disorder into a single scalar metric. Two patients with identical AHI values may present radically different clinical profiles, one dominated by severe nocturnal hypoxemia with minimal subjective symptoms, another characterized by pronounced daytime sleepiness with relatively modest desaturation [11]. Furthermore, the AHI fails to discriminate between pathophysiologically distinct event types: obstructive versus central events, events associated with severe versus mild desaturation, and events occurring predominantly in Rapid eye movement (REM) versus non REM sleep stages. As Zinchuk and Yaggi (2020) [4] argued in their comprehensive review, the AHI is analogous to using a single metric (e.g., blood pressure) to characterize the entire spectrum of cardiovascular disease, thereby obscuring the phenotypic diversity that is critical for precision management. This inadequacy has direct consequences for treatment selection: blanket therapeutic strategies based on AHI thresholds risk overtreating asymptomatic patients with elevated AHI values while undertreating those with lower AHI but significant cardiometabolic risk [16].

1.3 Phenotypic heterogeneity and the need for stratification

Over the past decade, a growing body of evidence has established that OSA is not a single, homogeneous entity but rather a syndrome encompassing multiple distinct phenotypes with divergent clinical presentations, comorbidity burdens, and therapeutic responses [4]. This recognition of phenotypic heterogeneity has been facilitated by the application of data driven clustering methodologies, which enable the identification of patient subgroups without imposing a prior assumptions about group membership [6; 17].

The landmark cluster analysis by Bailly et al. (2016) [16] of a French national OSA cohort of over 18,000 patients identified six distinct clinical phenotypes at the time of diagnosis, differentiated by constellations of symptoms, body mass index (BMI), age, and comorbidity patterns. Notably, some clusters were characterized by high daytime sleepiness and obesity (a “sleepy obese” phenotype), while others exhibited minimal subjective symptoms but carried a disproportionate cardiovascular burden (a “minimally symptomatic, cardiometabolic risk” phenotype). This study demonstrated that OSA is not reducible to a single severity continuum: the six cluster solution captured clinically actionable variation that AHI based severity categories alone would obscure. Similarly, Keenan et al. (2018) [17], in a multi center analysis spanning international sleep centers and thousands of patients, identified three recurrent clinical subtypes, namely disturbed sleep, minimally symptomatic, and excessively sleepy, reinforcing that certain phenotypic themes are robust across diverse populations.

The therapeutic implications of this heterogeneity are profound. Studies that have linked cluster derived phenotypes to treatment outcomes reveal that response to continuous positive airway pressure (CPAP) therapy is not uniform across patient subtypes. The “excessively sleepy” phenotype has consistently shown the greatest symptomatic benefit from CPAP, whereas the “insomnia like, disturbed sleep” phenotype derives comparatively less improvement [4]. Yassen et al. (2022) [18] using longitudinal data from the European Sleep Apnea Database (ESADA), demonstrated that baseline cluster membership predicts differential trajectories of treatment response, providing evidence that phenotyping may serve as a tool for treatment personalization rather than merely disease description. Similarly, Silveira et al. (2022) [7] assessed cluster derived phenotypes over a five year follow up and found that specific phenotypic groups exhibited prognostic significance in terms of long term outcomes, further supporting the clinical utility of phenotypic stratification.

Phenotypic heterogeneity in OSA is not limited to symptom and comorbidity profiles; polysomnographic data reveal additional dimensions of variability that further challenge the adequacy of the AHI as a unifying metric. Nakayama et al. (2019) [19] clustered male OSA patients based on respiratory event characteristics during polysomnography, identifying subgroups distinguished by the predominance of hypoxemia driven versus arousal driven events. Gasa et al. (2023) [20] similarly employed hierarchical clustering on PSG derived pathophysiological metrics and delineated three phenotypes with distinct patterns of oxygen desaturation, arousal, and positional dependency. These polysomnographic phenotypes illustrate that two patients with the same AHI can have fundamentally different pathophysiological expressions of the disorder, one primarily driven by intermittent hypoxemia with potential cardiovascular consequences, another by sleep fragmentation with predominantly neurocognitive effects [10].

Embracing phenotypic stratification, that is, classifying patients according to meaningful clinical attributes including symptom profiles, polysomnographic patterns, comorbidities, and demographic characteristics, offers a pathway toward more personalized clinical management, as conceptualized in Fig. 2. Rather than applying a uniform treatment algorithm to all patients above a given AHI threshold, phenotype informed approaches enable targeted interventions that match the specific risk profile and therapeutic needs of each subgroup [4; 6]. This transition from a one di-

mensional, AHI centric paradigm to a multidimensional, phenotype driven framework constitutes a necessary step toward precision medicine in sleep disordered breathing.

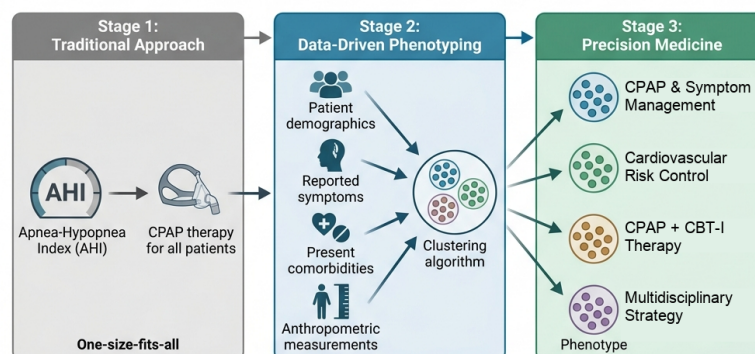


Figure 2: Conceptual framework for the transition from AHI-centric diagnosis to phenotype-driven precision medicine (Author's own elaboration).

1.4 Dissertation structure

This dissertation is organized in a sequential and progressive manner to address the proposed objectives. Chapter 1 introduces the clinical and socioeconomic context of OSA, articulates the inadequacy of the AHI as a sole diagnostic metric, and establishes the rationale for phenotypic stratification through unsupervised learning. Chapter 2 presents a comprehensive literature review on clustering methodologies applied to OSA phenotyping, identifying the principal methodological gaps and formulating the research objectives. Chapter 3 describes the study design, study populations, and the analytical framework, detailing the clinically guided and data driven approaches using the k-modes algorithm. Chapter 4 presents the experimental results obtained from both cohorts, detailing the baseline characteristics and the clinical phenotypes derived from each methodological approach. Chapter 5 provides the clinical interpretation and discussion of the identified phenotypes, contrasting the different variable selection strategies and contextualizing the findings within the broader literature. Finally, Chapter 6 synthesizes the overarching conclusions and outlines promising directions for future research in the field.

Chapter 2

Literature review

2.1 Introduction

The application of unsupervised learning to OSA phenotyping reflects a paradigm shift from the conventional AHI centric classification toward data driven patient stratification. As established in the preceding chapter, OSA is a heterogeneous condition whose clinical diversity is poorly captured by a single index. Cluster analysis, a family of statistical methods that partitions data points into groups of maximal internal similarity and maximal inter group dissimilarity [21], has emerged as the predominant unsupervised approach for identifying OSA phenotypes. A recent systematic review by Almeida et al. (2025) [22] identified 51 studies applying clustering to sleep disorders, of which 38 (74%) focused on OSA, underscoring the centrality of this condition in the phenotyping literature. The same review found that hierarchical clustering (20 studies, 39%) and k-means (14 studies, 27%) were the most frequently employed algorithms, followed by two step clustering (8 studies, 16%).

This chapter reviews the methodological trajectory of clustering in OSA research, progressing from traditional centroid based, hierarchical, and categorical methods (Section 2.2) to network based community detection approaches (Section 2.3). It then critically evaluates the limitations pervading the current literature (Section 2.4) and concludes by formulating the specific objectives of this dissertation (Section 2.5).

2.2 Clustering methods in OSA research

2.2.1 Overview of traditional algorithms

Early OSA phenotyping studies predominantly employed distance based partitioning algorithms, particularly k-means and agglomerative hierarchical clustering with Ward’s linkage [16; 17]. K-means operates by partitioning n observations into k clusters such that each observation belongs to the cluster whose centroid (arithmetic mean) minimizes the within cluster sum of squared Euclidean distances:

$$J = \sum_{j=1}^k \sum_{x_i \in C_j} \|x_i - \mu_j\|^2 \quad (2.1)$$

where C_j denotes the j th cluster, μ_j its centroid, and $\|\cdot\|$ the Euclidean norm [23]. The algorithm iterates between assigning observations to the nearest centroid and recomputing centroids until convergence. Hierarchical clustering, by contrast, constructs a nested hierarchy of merges (agglomerative) or splits (divisive), typically visualized as a dendrogram. Ward’s method, the

most commonly used linkage in OSA studies, minimizes the total within cluster variance at each merging step [16].

These algorithms were instrumental in establishing the foundational evidence that OSA encompasses multiple phenotypes rather than a single severity continuum. However, they operate under assumptions of continuous numerical data and spherical cluster geometry, which are often violated by clinical datasets composed predominantly of categorical and binary variables (e.g., presence/absence of comorbidities, symptom categories, sex) [6; 24].

2.2.2 K-modes clustering for categorical clinical data

A critical limitation of k-means and related centroid based methods is their reliance on Euclidean distance and arithmetic means, which are undefined or semantically meaningless for categorical variables [23]. Clinical OSA datasets frequently comprise binary (presence/absence of comorbidities) and nominal (symptom categories, sex) variables, rendering continuous distance assumptions inappropriate. The k-modes algorithm, proposed by Huang et al. (2023) [23], addresses this limitation by replacing the Euclidean distance with the *simple matching dissimilarity* measure and substituting cluster centroids (means) with cluster *modes* (most frequent category per variable).

Formally, for two categorical data points $X = (x_1, x_2, \dots, x_m)$ and $Y = (y_1, y_2, \dots, y_m)$, the simple matching dissimilarity is defined as:

$$d(X, Y) = \sum_{j=1}^m \delta(x_j, y_j), \quad \text{where} \quad \delta(x_j, y_j) = \begin{cases} 0 & \text{if } x_j = y_j \\ 1 & \text{if } x_j \neq y_j \end{cases} \quad (2.2)$$

This measure counts the number of mismatched categories between two observations, providing an intuitive and computationally efficient distance metric for categorical data. The k-modes algorithm proceeds analogously to k-means: (1) initialize *k*-modes (either randomly or via a frequency based heuristic); (2) assign each observation to the cluster whose mode minimizes the dissimilarity; (3) update modes by computing the most frequent category for each variable within each cluster; (4) iterate until convergence [23].

Ferreira-Santos and Rodrigues pioneered the application of k-modes to OSA phenotyping in a series of studies. In their initial work, they applied k-modes to a Portuguese cohort using demographic, comorbidity, and symptom variables, demonstrating the feasibility of categorical clustering for identifying clinically interpretable OSA subgroups [24; 25]. In a subsequent study, Ferreira-Santos and Rodrigues (2023) [26] applied k-modes to 264 patients suspected of having OSA and identified three phenotypic clusters based exclusively on pre-triage clinical variables, without incorporating PSG derived data. This approach eliminated incorporation bias, a pervasive methodological flaw wherein post diagnostic variables are used to construct the phenotyping model, thereby precluding its use as a pre test triage tool [22]. In a parallel study, Ferreira-Santos and Rodrigues (2021) [6] developed a complete phenotyping algorithm that was validated across independent cohorts, reporting an Area Under the Curve (AUC) of 72 to 84%, sensitivity of 93%,

specificity of 38%, and a diagnostic odds ratio of up to 8.14. This constitutes the only study in the literature to combine pre diagnostic clustering with rigorous external validation.

The k-modes algorithm is now recognized as an established method for handling categorical clinical data in OSA phenotyping, and its adoption reflects a growing awareness that algorithmic choice must be aligned with the nature of the input variables [6; 26].

2.2.3 Phenotypic findings from traditional clustering

Traditional clustering has consistently yielded between 3 and 6 phenotypic clusters, revealing recurrent themes across studies. The landmark analysis by Bailly et al. (2016) [16] on a French national cohort of 18,263 diagnosed OSA patients identified six clusters using hierarchical clustering with Ward's method. These included: (1) a young, non obese, minimally symptomatic group; (2) an elderly cluster with high cardiovascular comorbidity burden but few subjective symptoms; (3) an obese, sleepy phenotype with metabolic syndrome; and (4 to 6) intermediate groups varying in symptom severity and comorbidity patterns. The identification of a "minimally symptomatic but high risk" phenotype was clinically significant, as it demonstrated that patients with few subjective complaints can nevertheless carry a substantial cardiometabolic burden that would be missed by symptom based screening.

Keenan et al. (2018) [17], in a multi center international study spanning 13 sleep centers and 4,603 participants, applied k-means clustering and identified three recurrent subtypes: disturbed sleep, minimally symptomatic, and excessively sleepy. The consistency of these subtypes across geographically and ethnically diverse populations supported the robustness of symptom driven phenotyping. However, the clusters were primarily driven by symptom questionnaire scores, with limited integration of comorbidity or physiological data.

At the polysomnographic level, several studies have demonstrated that patients sharing identical AHI values can be clustered into physiologically distinct subgroups. Nakayama et al. (2019) [19] identified three PSG based phenotypes in male OSA patients distinguished by the predominance of hypoxemia driven versus arousal driven respiratory events. Gasa et al. (2023) [20] similarly employed divisive hierarchical clustering with Gower's distance on PSG derived pathophysiological metrics and delineated three phenotypes: predominantly hypoxemic, predominantly arousal driven, and a mixed pattern. Zinchuk et al. (2018) [11] applied k-means clustering to polysomnographic features in 1,247 OSA patients and identified seven phenotypes, demonstrating that certain PSG defined subtypes carry significantly elevated cardiovascular risk independent of AHI severity. Kim et al. (2020) [10] confirmed these findings in a Korean cohort, showing that PSG derived phenotypes have differential implications for mortality.

Beyond symptom and PSG based clustering, several studies have focused on specific clinical dimensions. Joosten et al. (2012) [27] clustered 1,064 patients with mild to moderate OSA using k-means and identified six phenotypes differentiated by BMI, age, sex, and Epworth Sleepiness Scale scores. Romero-Peralta et al. (2022) [28] applied hierarchical clustering to 1,655 Spanish OSA patients and identified four phenotypes with implications for patient management. Zhang et al. (2022) [29] performed a multidimensional cluster analysis integrating clinical, demographic,

and PSG variables in 291 patients and obtained three clusters. Venkatnarayan et al. (2023) [30] used two step clustering in 100 patients and similarly identified three phenotypes. Wei et al. (2023) [31] clustered 1,118 patients based on sleep perception and quality measures and found five distinct groups. These results collectively confirm that the heterogeneity of OSA manifests across multiple axes, including symptom burden, physiological derangement, comorbidity patterns, and demographic characteristics, and that no single axis is sufficient to capture the full spectrum of phenotypic diversity.

Anatomical and craniofacial dimensions have also been explored. Huang et al. (2023) [32] applied k-means clustering to cephalometric variables in 722 OSA patients and identified three anatomical phenotypes with distinct patterns of upper airway obstruction. An et al. (2020) [33] extended this perspective by applying k-means to a smaller cohort of 89 patients, integrating OSA severity, obesity, and craniofacial pattern into a combined clustering framework. Kim et al. (2020) [34] performed multi perspective clustering integrating craniofacial, PSG, and clinical variables in 421 patients, identifying three phenotypes relevant to surgical versus non surgical therapeutic decisions.

Comorbidity focused analyses have further enriched the phenotypic landscape. Vavougiou et al. (2016) [35] employed categorical principal component analysis combined with two step clustering in 1,472 patients and identified six comorbidity based phenotypes. Ntenta et al. (2022) [36] applied two step clustering in a primary care population and similarly identified comorbidity driven clusters. Testelmans et al. (2022) [37] used self organizing maps to cluster 710 patients with moderate to severe OSA by comorbidity profiles, identifying nine distinct clusters that differed significantly in cardiovascular and metabolic burden.

While the studies reviewed so far have focused on phenotype identification at a single time point, a growing number of investigations have begun to integrate longitudinal outcome data, shifting the emphasis from phenotype discovery toward clinical validation and prognostic utility. Silveira et al. (2022) [7] assessed cluster derived phenotypes at 5 year follow up and found prognostic significance. Yassen et al. (2022) [18] linked baseline clusters to CPAP treatment response trajectories using ESADA longitudinal data. Tondo et al. (2024) [38] combined hierarchical clustering with machine learning classifiers to explore phenotypes for long term mortality stratification, reporting AUC values of 0.947 to 0.948. Hamada et al. (2024) [39] used hierarchical clustering to predict CPAP adherence patterns. These studies represent a critical evolution from phenotype *discovery* toward phenotype *utility*.

Advanced algorithmic approaches have also been explored, though less frequently. Ma et al. (2021) [40] applied a Gaussian Mixture Model with a Dirichlet Process prior to 2,277 OSA patients and identified six clusters, subsequently combining unsupervised and supervised learning. Mai et al. (2018) [41] developed an evolutionary active constrained clustering algorithm specifically designed for OSA analysis in 22,568 patients. Lacedonia et al. (2016) [42] employed correlation network clustering in 198 patients. Fontanilles-Arbones et al. (2024) [3] applied divisive hierarchical clustering with Gower's distance to a female specific cohort of 1,886 patients, identifying four phenotypes and highlighting sex specific phenotypic patterns. Oki et al. (2024)

[43] and Koza et al. (2021) [12] investigated the relationship between cluster membership and biomarkers such as leptin levels. Ida et al. (2022) [44] identified five unique clinical phenotypes in a Japanese population. Kim et al. (2021) [45] examined the correlation between OSA subtypes and mood states. Dai et al. (2023) [46] evaluated a novel sleep breathing impairment index using two step clustering. Tsuchiya et al. (1992) [47], in one of the earliest cluster analyses of OSA, identified two craniofacial subtypes using hierarchical clustering.

2.2.4 Comparative algorithm evaluation

Despite the extensive application of clustering to OSA, surprisingly few studies have systematically compared multiple algorithms on identical datasets. Ghorvei et al. (2025) [48] conducted one of the first rigorous comparative analyses, applying k-means, agglomerative hierarchical clustering, fuzzy c means, and Gaussian mixture models to the same set of PSG derived variables in 865 patients. They evaluated cluster quality using the Calinski Harabasz Index (CHI), Davies Bouldin Index (DBI), Dunn Index (DI), and silhouette coefficient. The results revealed that the choice of algorithm significantly influenced the resulting phenotypic partition: k-means yielded a CHI of 102.037 and a silhouette coefficient of 0.090, while hierarchical clustering produced a lower silhouette coefficient of 0.062 and a DBI of 2.673. No single algorithm was universally optimal across all validity indices, underscoring the risk of method dependent results and the need for multi method validation in OSA phenotyping studies.

2.3 Emerging methods: network based and multimodal clustering

A fundamentally different approach to phenotyping involves the application of network science and graph based community detection algorithms. In this framework, patients are represented as nodes in a graph, and edges between nodes encode pairwise clinical similarity. Clusters, termed “communities” in network terminology, are then identified by detecting dense subnetworks of patients who are more similar to one another than to members of other communities [49; 50].

The key theoretical construct underlying this approach is *modularity*, a scalar quantity that measures the fraction of edges falling within communities minus the expected fraction if edges were distributed randomly [51]. Formally, the modularity Q of a partition is defined as:

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \quad (2.3)$$

where A_{ij} is the weight of the edge between nodes i and j , $k_i = \sum_j A_{ij}$ is the sum of weights attached to node i , $m = \frac{1}{2} \sum_{i,j} A_{ij}$ is the total edge weight, and $\delta(c_i, c_j) = 1$ if nodes i and j belong to the same community and 0 otherwise. The Louvain algorithm, proposed by Blondel et al. (2008) [51], is the most widely used method for modularity optimization due to its computational efficiency on large networks. It operates in two iterative phases: (1) each node is initially assigned to its own community, and nodes are moved to neighboring communities if doing so increases Q ; (2) a new meta network is constructed in which nodes represent communities from the previous phase,

and the process repeats until no further improvement in Q is achieved. A key advantage is that the number of communities is determined endogenously by the optimization process, eliminating the need to pre specify k [50].

The application of network based clustering to OSA was pioneered by Mihaicuta et al. (2017) [49], who constructed patient similarity networks from clinical and PSG variables in 1,371 patients and applied the Girvan Newman algorithm for community detection, identifying eight communities. They demonstrated that network metrics could predict OSA severity and meaningfully subgroup patients. Topîrceanu et al. (2020) [52] extended this work by constructing patient networks from multiple clinical and physiological variables and applying community detection to explore gender specific phenotypic differences, finding that network analysis identified distinct male and female phenotype communities. Lacedonia et al. (2016) [42] employed correlation network clustering in 198 patients and identified three communities.

A critical advancement was achieved by Gutiérrez-Tobal et al. (2021) [53], who developed *subject based weighted correlation networks* for pediatric sleep apnea phenotyping. In this approach, edges between patient nodes are weighted by the correlation of their SpO₂ signal features, and modularity optimization (via the Louvain algorithm) is used to identify communities. The study demonstrated that this network based framework could capture multidimensional disease structure and reveal subtle phenotypic markers in children that conventional clustering methods overlook. Gómez-Pilar et al. (2022) [50] further refined this methodology by applying subject based association networks to pediatric OSA, using the ForceAtlas2 layout algorithm for network visualization and the Louvain algorithm for community detection [54].

Most recently, Gómez-Pilar et al. (2025)[55] applied SpO₂ based weighted correlation networks to an adult OSA cohort, directly comparing the communities detected via modularity optimization against the clusters derived from k-means on the same dataset. The network based approach identified subgroups that captured significant differences in seven anthropometric, cardiac, and sleep related features that k-means failed to detect, providing the first direct evidence that association networks can outperform traditional clustering in adult OSA phenotyping.

Despite these promising results, association networks remain a fundamental research gap in adult OSA populations. The pediatric applications by Gutiérrez-Tobal et al. (2021) [53] and Gómez-Pilar et al. (2022) [50] demonstrated the methodology's potential, but the translation to adult cohorts using pre diagnostic clinical variables, rather than PSG derived oximetry features, has not yet been systematically undertaken. This gap between pediatric and adult research, alongside the historical underutilization of categorical clustering algorithms, constitutes a key methodological opportunity in the field. While the present dissertation focuses on k-modes clustering to address the categorical nature of pre-triage clinical data, the potential of association networks for adult OSA populations is highlighted as a promising parallel direction in the ongoing evolution of phenotyping methodologies.

2.4 Limitations in current research

Despite the considerable volume of published work, OSA phenotyping literature is characterized by several critical and pervasive limitations that constrain both scientific rigor and clinical translatability. The most consequential limitation is the overwhelming dependence on post diagnostic, polysomnography derived variables. As demonstrated by Almeida et al. (2025) [22], the majority of studies incorporate such measures as clustering inputs, introducing a profound incorporation bias that fundamentally precludes the use of these models as independent pre test stratification tools, since the variables driving the classification are only available after the patient has already undergone the diagnostic procedure. Furthermore, this limitation is compounded by a pervasive absence of validation; most studies fail to report internal validation metrics, and external validation across independent cohorts remains exceptionally rare, as only Ferreira-Santos and Rodrigues (2021) [6] have rigorously performed it. Without such validation, it is impossible to ascertain whether the reported phenotypes represent stable patient groupings or mere statistical artefacts, a concern amplified by Ghorvei et al. (2025) [48], who showed that different algorithms yield substantially different cluster solutions on the same dataset. Another critical issue is the algorithm data mismatch, where continuous distance algorithms like k-means are inappropriately applied to categorical clinical variables, distorting the feature space, whereas methods specifically designed for such data, such as k-modes, remain underutilized [23; 26]. Additionally, significant demographic and methodological gaps persist, notably the failure to translate advanced network based community detection methods, which have been successfully deployed in pediatric cohorts by Gutiérrez-Tobal et al. (2021 [53] and Gómez-Pilar et al. (2022) [50], to adult populations using pre diagnostic data. Finally, the practical value of most reported phenotypes remains largely speculative due to a distinct lack of linkage to clinically meaningful outcomes, such as disease progression, treatment response, or long term mortality, with only a few notable exceptions attempting to bridge this gap [18; 7].

2.5 Objectives

This dissertation is guided by fundamental research questions aimed at understanding whether phenotyping based exclusively on pre-diagnostic clinical and demographic data can identify meaningful and distinct OSA patient subgroups. Specifically, this work seeks to explore how different variable selection strategies and population heterogeneities impact the clustering outcomes. The primary research questions include:

- Can phenotyping based exclusively on pre-triage clinical and demographic data identify meaningful and distinct OSA patient subgroups?
- How do different variable selection strategies, specifically, a clinically guided approach versus a fully data-driven approach affect the resulting k-modes cluster solutions and the clinical interpretability of OSA phenotypes?

- What are the baseline epidemiological and clinical differences between the Portuguese and Spanish hospital cohorts, and how does this population heterogeneity influence the derived clustering models?

Based on these inquiries, the primary goal of this dissertation is to identify and describe clinically meaningful OSA phenotypes using the k-modes clustering algorithm applied to pre-triage categorical data. To achieve this, the work is structured around the following specific objectives:

- To curate, harmonize, and integrate clinical, demographic, and anthropometric data from patients recruited across the Portuguese and Spanish cohorts, systematically evaluating their baseline differences.
- To apply the k-modes clustering algorithm using two complementary analytical strategies: a clinically guided approach (based on univariate feature selection) and a data-driven approach (incorporating all available clinical variables).
- To compare the clustering results derived from these two approaches, evaluating their impact on cluster granularity, stability, and clinical utility.
- To interpret the resulting clusters by rigorously analyzing differences in symptom profiles, comorbidity patterns, and demographic characteristics across both cohorts, discussing the implications of dataset heterogeneity on personalized OSA diagnosis and management.

Chapter 3

Methods

This chapter details the methodological framework designed to identify and characterize clinically meaningful OSA phenotypes. The study adopts a structured analytical pipeline divided into two complementary approaches: a clinically guided approach and a data driven approach allowing for a comprehensive exploration of patient data from distinct but converging perspectives.

3.1 Study design and population

This is a retrospective, multicenter, cross sectional study based on previously collected clinical data obtained from two independent tertiary hospitals: Unidade Local de Saúde de São João (ULSSJoão), Porto (Portugal) and Hospital Universitario Río Hortega, Valladolid (Spain). All data were acquired during routine clinical sleep studies in adult patients (≥ 18 years) referred for suspected OSA. Data collection adhered to standard clinical protocols at both institutions and followed ethical guidelines for the use of retrospective data, with all patient information fully anonymized prior to analysis. This study was approved by the Ethics Committee of Unidade Local de Saúde de São João (approval no. 190/2024).

To account for population differences and ensure methodological robustness, the cluster analysis was conducted separately for the Portuguese and Spanish datasets; the two populations were not merged into a single joint analysis. The Portuguese cohort comprises a total of 188 adult patients evaluated at ULSSJoão using different levels of polysomnography, specifically including in-lab polysomnography (PSG), Level II, and home sleep apnea testing (HSAT) studies. The Spanish cohort consists of 355 adult patients evaluated using in-lab PSG at Hospital Universitario Río Hortega.

In both cohorts, sleep stages and respiratory events, including apneas and hypopneas, were scored according to the 2012 American Academy of Sleep Medicine (AASM) criteria [56]. Although PSG was used to establish the clinical diagnosis and severity classification of OSA (via the AHI), the primary focus of the subsequent analytical approaches remains grounded in pre test clinical and demographic variables, aiming to phenotype patients before resorting to complex polysomnographic features.

3.2 Variables and data description

Prior to the main analyses, a comprehensive descriptive assessment was conducted to characterize the study populations. A total of 23 pre-triage clinical variables were collected and harmonized across both datasets. To achieve this harmonization, a unified data dictionary was developed to integrate both cohorts. This dictionary served as a standardized reference to map clinical features

to a common nomenclature, standardize coding schemes for comorbidities and symptoms (e.g., reconciling ‘yes/no’ formats with ‘1/0’ numeric codes), and resolve divergent categorizations for multi-class variables such as tobacco use (where the Portuguese cohort used ‘yes/no/ex’ and the Spanish used ‘1/0/2’). To ensure the accurate interpretation of these numeric values, the original institutional codebooks were consulted and local clinical teams were engaged to confirm that, for example, the numeric code ‘2’ in the Spanish dataset correctly corresponded to the ‘*ex-smoker*’ category. By resolving these structural and linguistic discrepancies, the harmonization process ensured rigorous data comparability between the two tertiary hospitals. Crucially, the study employed exclusively categorical variables throughout the analytical pipeline. Variables that were originally recorded on a continuous scale were systematically categorized into clinically meaningful discrete groups prior to analysis, as detailed below.

3.2.1 Variable categorization

Three originally continuous variables were converted into categorical form according to established clinical thresholds. Age was categorized into three strata: 20–44 years, 45–65 years, and 66–90 years, reflecting clinically distinct phases of OSA presentation and comorbidity accumulation [16]. Body Mass Index (BMI) was categorized according to the World Health Organization (WHO) classification into five groups: Underweight/Normal weight ($\text{BMI} < 25 \text{ kg/m}^2$), Overweight (25–29.9), Obesity Grade I (30–34.9), Obesity Grade II (35–39.9), and Obesity Grade III (≥ 40). The Epworth Sleepiness Scale (ESS) was dichotomized into Normal ($\text{ESS} < 10$) and Excessive Daytime Sleepiness (EDS; $\text{ESS} \geq 10$), based on the standard clinical interpretation threshold [6].

3.2.2 Variable inventory

The full set of variables is organized into four domains. The demographic domain comprises three variables: sex (male/female), age (three strata), and BMI (five categories). The behavioral and lifestyle domain includes three variables: alcohol consumption (yes/no), tobacco use (yes/no), and history of bariatric surgery (yes/no). The comorbidity domain encompasses 13 binary variables (present/absent): depression, arterial hypertension, cerebrovascular disease, asthma, chronic obstructive pulmonary disease (COPD), diabetes, dyslipidemia, thyroid disease, dementia, cancer, gastroesophageal reflux, and glaucoma. Finally, the symptom domain includes four variables: Epworth Sleepiness Scale (two categories), gasping (yes/no), snoring (yes/no), witnessed apneas (yes/no), and insomnia (yes/no).

This unified categorical format ensures compatibility with the k-modes clustering algorithm, which computes dissimilarity via the simple matching coefficient rather than Euclidean distance [23].

3.3 Analytical approaches

To thoroughly investigate the clinical data and uncover underlying patient phenotypes, the methodology was structured around two complementary analytical approaches. Rather than representing a distinction between different classes of algorithms, the two approaches differ in their variable selection philosophy: Approach 1 employs a *clinically guided* strategy that filters variables based on their individual association with the diagnostic outcome, whereas Approach 2 adopts a *data driven* strategy that includes the full variable set without reference to the diagnostic label. Both approaches apply the same core clustering algorithm (k-modes) to the selected variable subsets.

3.3.1 Approach 1: clinically guided feature selection

The first approach is clinically oriented, focusing on the identification and selection of pre-triage variables that exhibit a statistically significant association with OSA status. This strategy aims to reduce dimensionality and isolate the most clinically relevant predictors before clustering, effectively filtering out background noise from non discriminative variables.

3.3.1.1 Univariate analysis and feature selection

A univariate comparative analysis was performed to evaluate the relationship between each of the 23 categorical variables and the diagnostic outcome (OSA vs. healthy). Two statistical tests were employed, selected according to the distributional properties of the data:

- **Pearson’s Chi Square test (χ^2):** applied when the expected frequency in each cell of the contingency table was ≥ 5 , as required by the large sample assumption underlying the χ^2 distribution [57].
- **Fisher’s Exact test:** applied when one or more cells had an expected frequency < 5 , as the χ^2 approximation becomes unreliable with sparse data. Fisher’s test computes the exact probability of obtaining the observed (or a more extreme) distribution under the null hypothesis of independence, and is therefore appropriate for small sample or low frequency comparisons [57].

Following this analysis, a predefined significance threshold of $p < 0.05$ was adopted for variable retention. Only variables demonstrating a statistically significant association with OSA status were selected for inclusion in the subsequent clustering step.

3.3.1.2 Cluster analysis

The k-modes algorithm [23] was applied to the selected variable subset. To determine the optimal number of clusters (k), two complementary methods were employed:

- **Elbow method:** the clustering cost (total within cluster dissimilarity) was computed for a range of k values (typically $k = 2$ to $k = 10$), and the point at which the rate of cost reduction diminishes (the “elbow”) was identified as the candidate optimal k .

- **Silhouette method:** the average silhouette coefficient was computed for each candidate k , measuring how similar each observation is to its own cluster relative to the nearest neighboring cluster. Higher silhouette values indicate more cohesive and well separated clusters.

The two methods were used in conjunction: the elbow method provided an initial candidate range for k , and the silhouette method was used to confirm or refine the selection within that range. Once the optimal k was determined, the final k-modes model was fitted with multiple random initializations to ensure stability and reproducibility of the cluster assignments.

A post hoc analysis was then performed, cross referencing the identified clusters with the AHI based clinical diagnosis and severity classification (normal, mild, moderate, severe). By retaining AHI as a validation metric rather than an input variable, this approach avoids the circularity that arises when severity indices are used simultaneously for cluster formation and cluster evaluation.

3.3.2 Approach 2: data driven full variable inclusion

The second approach provides a data driven perspective, following the clustering framework recommended by the Spanish data analysis team. In contrast to Approach 1, this strategy does not perform any prior filtering based on the diagnostic label; instead, all 23 pre-triage clinical variables are included as input features for the k-modes algorithm. No information about whether a patient has OSA or not is provided to the model during cluster construction.

The primary objective of this approach is to identify natural, underlying patterns and distinct phenotypic subgroups within the patient cohorts based on the full spectrum of available clinical information, including behavioral, comorbid, and symptomatic variables that may not individually reach statistical significance in a univariate analysis but may contribute to multivariate phenotypic structure.

3.3.2.1 Cluster analysis

The same analytical pipeline described in Section 3.3.1.2 was employed: the optimal number of clusters (k) was determined via the combined use of the elbow and silhouette methods, followed by final k-modes model construction with multiple random initializations.

Once the clustering solution was obtained, a subsequent post hoc analysis was performed, cross referencing the data driven clusters with the clinical diagnosis and AHI severity. Any phenotypic structure that emerges under this approach and is also observed under Approach 1 can be considered more robust, as it has been replicated across two independent variable selection strategies.

3.3.3 Software and implementation

All analyses were implemented in Python (version 3.x). The k-modes algorithm, originally proposed by Huang [23], was executed using the `kmodes` Python library [58]. Descriptive statistics, contingency table analyses (chi-square and Fisher's exact tests), and visualization were performed using the `pandas`, `scipy`, and `matplotlib` libraries.

3.3.3.1 Visualization strategies

To facilitate clinical interpretation of the clustering results, two complementary visualization strategies were employed using the `matplotlib` library: heatmaps and radar plots.

Heatmaps encode variable prevalence as color intensity within a two-dimensional matrix, excelling at representing the absolute distribution density across all clusters simultaneously. This format enables rapid identification of variables that are universally present versus those that sharply discriminate between subgroups.

Because heatmaps display variables independently and lack geometric context, radar plots were additionally utilized to capture the multivariate “morphotype” of each cluster. By mapping a cluster’s profile as a polygon across normalized axes, the area and symmetry of the shape provide immediate visual cues regarding clinical burden. To make these visual cues clinically intuitive, the variables on the radar plot were explicitly grouped into three distinct domains: metabolic (BMI categories), cardiovascular (cardiometabolic comorbidities), and symptomatic (patient-reported symptoms). This deliberate grouping ensures that the direction in which the polygon expands immediately reveals the dominant clinical pathway (metabolic, cardiovascular, or symptomatic) characterizing that specific cluster.

Visualization strategies were then tailored to each analytical approach to maximize interpretability. Under the clinically guided approach (Approach 1), both heatmaps and radar plots were constructed exclusively with the subset of variables that demonstrated statistical significance in the univariate analysis.

Conversely, under the data-driven approach (Approach 2), heatmaps were generated using all 23 clinical variables to provide a comprehensive overview. However, to prevent excessive visual clutter and facilitate cross-cohort comparison, the corresponding radar plots were constructed using only the most discriminative subset of variables.

Chapter 4

Results

This chapter presents the results of the clustering analyses performed on the Portuguese and Spanish cohorts. The analyses were conducted independently for each dataset, applying two complementary variable selection strategies: a clinical approach (Section 4.1), guided by domain expertise, and a scientific approach (Section 4.2), utilizing the full set of available variables. For each approach, baseline cohort characteristics, cluster composition, and modal clinical profiles are reported. Additionally, two complementary visualization strategies: heatmaps and radar plots are employed to represent the identified cluster profiles.

4.1 Clinical approach

Following the clinically guided variable selection strategy described in Section 3.3.1, the k-modes algorithm was applied independently to the Portuguese and Spanish datasets. The results for each cohort are presented below.

4.1.1 Portuguese dataset

4.1.1.1 Baseline characteristics

Of the 188 patients included in the Portuguese cohort, 110 (58.5%) had a confirmed diagnosis of OSA, whereas 78 (41.5%) were classified as healthy. Table 1 describes the baseline characteristics of both groups. The proportion of males was higher in the OSA group (65/110, 59.1%) compared to the healthy group (37/78, 47.4%), although this difference did not reach statistical significance ($p = 0.152$). The age distribution was comparable across groups ($p = 0.341$), with the majority of patients in both groups falling within the 45–65 age stratum (53.8% healthy; 51.8% OSA).

A statistically significant difference was found in BMI distribution ($p = 0.004$): underweight or normal weight patients were more prevalent in the healthy group (19/78, 24.4%) than in the OSA group (8/110, 7.3%), while Obesity Grade I was more common in the OSA group (39/110, 35.5% vs. 18/78, 23.1%). Regarding comorbidities, arterial hypertension was significantly more prevalent among OSA patients (68/110, 61.8% vs. 35/78, 44.9%; $p = 0.031$). Insomnia was reported by a significantly higher proportion of OSA patients (62/110, 56.4%) compared to healthy subjects (27/78, 34.6%; $p = 0.005$). Snoring was nearly universal in both groups (97.4% healthy; 100% OSA), and the Epworth Sleepiness Scale showed no significant difference between groups ($p = 0.885$), with the majority classified as normal sleepiness in both populations.

The remaining comorbidities, including diabetes ($p = 0.539$), dyslipidemia ($p = 0.120$), cerebrovascular disease ($p = 0.169$), and COPD ($p = 0.207$), did not differ significantly between groups, although trends toward higher prevalence in the OSA group were observed for several of these variables.

Table 1: Descriptive characteristics of the Portuguese study cohort, n (%). P-values obtained using the chi square test unless otherwise specified.

Variable	Healthy (n=78)	OSA (n=110)	p value
Male gender	37 (47.4)	65 (59.1)	0.152
Age			0.341
20–44	14 (17.9)	13 (11.8)	
45–65	42 (53.8)	57 (51.8)	
66–90	22 (28.2)	40 (36.4)	
BMI (kg/m²)			0.004
Underweight / Normal	19 (24.4)	8 (7.3)	
Overweight	36 (46.2)	47 (42.7)	
Obesity Grade I	18 (23.1)	39 (35.5)	
Obesity Grade II	5 (6.4)	13 (11.8)	
Obesity Grade III	0 (0)	3 (2.7)	
Alcohol	34 (43.6)	59 (53.6)	0.226
Tobacco	11 (14.1)	25 (22.7)	0.196
Bariatric surgery	4 (5.1)	2 (1.8)	0.235 [#]
Depression	12 (15.4)	24 (21.8)	0.359
Hypertension	35 (44.9)	68 (61.8)	0.031
Cerebrovascular	6 (7.7)	17 (15.5)	0.169
Asthma	11 (14.1)	13 (11.8)	0.810
COPD	10 (12.8)	7 (6.4)	0.207
Diabetes	16 (20.5)	28 (25.5)	0.539
Dyslipidemia	37 (47.4)	66 (60.0)	0.120
Thyroid disease	10 (12.8)	16 (14.5)	0.902
Dementia	1 (1.3)	0 (0)	0.415 [#]
Cancer	11 (14.1)	15 (13.6)	>0.999 [#]
GE reflux	3 (3.8)	2 (1.8)	0.651 [#]
Glaucoma	1 (1.3)	2 (1.8)	>0.999 [#]
Epworth (ESS)			0.885
Normal	66 (84.6)	91 (82.7)	
EDS	12 (15.4)	19 (17.3)	
Gasping	7 (9.0)	7 (6.4)	0.697
Snoring	76 (97.4)	110 (100.0)	0.171 [#]
Witnessed apneas	67 (85.9)	99 (90.0)	0.527
Insomnia	27 (34.6)	62 (56.4)	0.005

[#] Fisher exact test. Abbreviations: BMI, body mass index; COPD, chronic obstructive pulmonary disease; ESS, Epworth Sleepiness Scale; EDS, excessive daytime sleepiness; GE, gastroesophageal. Significant p values ($p < 0.05$) are shown in bold.

4.1.1.2 Cluster characterization

Using the variables identified as clinically relevant, the k-modes algorithm was applied to the full Portuguese cohort ($n = 188$). The optimal number of clusters was determined via the elbow method and silhouette method, yielding a six cluster solution. Table 2 presents the detailed clinical characteristics of each cluster, stratified by the three primary discriminating variables: hypertension, insomnia, and BMI category.

The six clusters exhibited a clear combinatorial structure across these three axes. Clusters 1 and 2, the two largest groups ($n = 42$ and $n = 43$, respectively), were both characterized by a

predominance of obesity Grade I (64.3% and 69.8%, respectively), with the majority of patients presenting hypertension (71.4% and 67.4%). The distinguishing feature between these two clusters was insomnia: all patients in Cluster 1 reported insomnia (100%), whereas no patient in Cluster 2 did (0%). This binary split suggests that insomnia functions as an independent stratification variable within the obese, hypertensive subpopulation.

Clusters 4 and 6, although smaller ($n = 17$ and $n = 27$), shared universal hypertension (100%) and an exclusively overweight BMI profile. Again, they diverged on insomnia: Cluster 4 included exclusively insomniac patients (100%), whereas Cluster 6 included none (0%). Clusters 3 and 5 constituted the normotensive groups, once more splitting along the insomnia axis: Cluster 3 ($n = 30$) comprised normotensive patients with universal insomnia, predominantly overweight (60.0%), while Cluster 5 ($n = 29$) included normotensive, non insomniac patients, also predominantly overweight (72.4%).

Table 2: Clinical characteristics of the Portuguese cohort by the defined clusters, n (%).

Variable	C1 (n=42)	C2 (n=43)	C3 (n=30)	C4 (n=17)	C5 (n=29)	C6 (n=27)
Hypertension	30 (71.4)	29 (67.4)	0 (0)	17 (100)	0 (0)	27 (100)
Insomnia	42 (100)	0 (0)	30 (100)	17 (100)	0 (0)	0 (0)
BMI						
UW / Normal	7 (16.7)	8 (18.6)	7 (23.3)	0 (0)	5 (17.2)	0 (0)
Overweight	0 (0)	0 (0)	18 (60.0)	17 (100)	21 (72.4)	27 (100)
Obesity I	27 (64.3)	30 (69.8)	0 (0)	0 (0)	0 (0)	0 (0)
Obesity II	7 (16.7)	4 (9.3)	4 (13.3)	0 (0)	3 (10.3)	0 (0)
Obesity III	1 (2.4)	1 (2.3)	1 (3.3)	0 (0)	0 (0)	0 (0)

Abbreviations: BMI, body mass index; UW, underweight; C, cluster.

4.1.1.3 Cluster profiles and phenotype definition

On the basis of the cluster characterization described above, Table 3 summarizes the modal clinical profiles of each cluster. The AHI values, although not used as input variables for clustering, were computed post hoc for each cluster to assess the relationship between the identified phenotypes and disease severity.

Cluster 1 (obese, hypertensive, insomniac) exhibited the highest mean AHI (33.3 ± 22.3) with a modal severity of severe, suggesting that the concurrent presence of obesity, hypertension, and insomnia is associated with the most severe expression of OSA. In contrast, Cluster 5 (overweight, normotensive, non insomniac) presented the lowest mean AHI (14.9 ± 16.9) with a modal severity of mild, consistent with a low risk phenotypic profile. Notably, the hypertensive overweight clusters (C4 and C6) presented intermediate AHI values (29.9 and 16.6, respectively), with the insomniac subgroup (C4) exhibiting a higher burden than the non insomniac subgroup (C6), reinforcing the independent contribution of insomnia to OSA severity within this BMI stratum.

Table 3: Modes of the clustered clinical presentations of OSA – Portuguese cohort (clinical approach).

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
<i>Demographics</i>						
BMI	Obesity I	Obesity I	Overweight	Overweight	Overweight	Overweight
<i>Comorbidities</i>						
Hypertension	Yes	Yes	No	Yes	No	Yes
<i>Symptoms</i>						
Insomnia	Yes	No	Yes	Yes	No	No
<i>AHI</i>						
Mean (SD)	33.3 (22.3)	23.0 (18.1)	25.5 (23.7)	29.9 (14.5)	14.9 (16.9)	16.6 (14.5)
Median (IQR)	30.0	18.2	16.8	26.8	9.8	12.9
Severity	Severe	Moderate	Mild	Moderate	Mild	Moderate

Modes represent the most frequent category per variable within each cluster. AHI severity (mode) is presented alongside the mean and median AHI values. AHI was not used as a clustering variable.

4.1.1.4 Visualization of cluster profiles

To facilitate the interpretation of the six-cluster solution, two complementary visualization strategies were employed: a heatmap encoding absolute variable prevalence, and a radar plot encoding the multivariate morphotype of each cluster.

The heatmap in Fig. 3 provides a visual summary of the variable distributions across the six clusters, highlighting the sharp boundaries between clusters on hypertension, insomnia, and BMI. Each cell encodes the prevalence of a given variable within a given cluster as color intensity, enabling rapid identification of variables that are universally present (e.g., overweight) versus those that sharply discriminate between clusters (e.g., hypertension, insomnia).

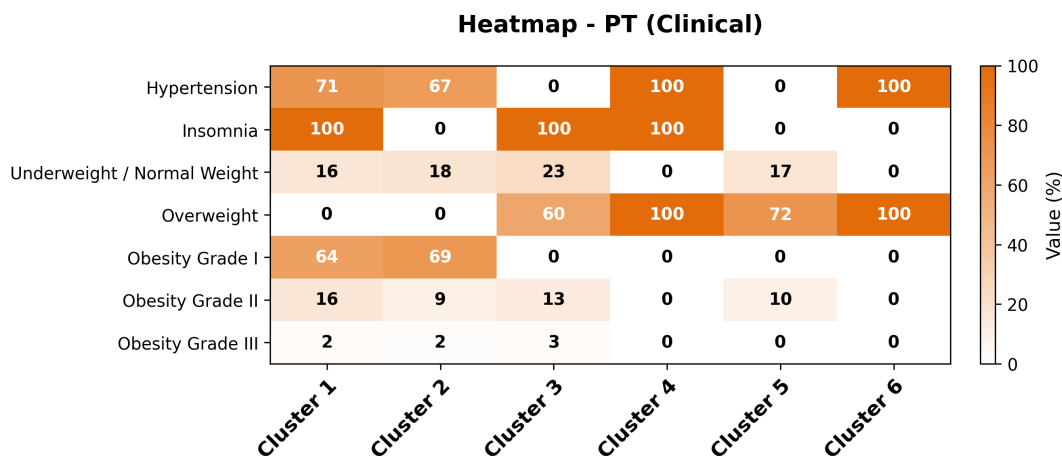


Figure 3: Heatmap of categorical variable distributions across the six clusters identified in the Portuguese cohort (clinical approach).

The radar plot in Fig. 4 maps each cluster's profile as a polygon whose vertices correspond to

the three principal discriminating axes: hypertension, insomnia, and the BMI categories (underweight/normal, overweight, Obesity I, II, and III). The six polygons exhibit distinct morphotypes. Clusters 1 and 2 (the obese hypertensive pair) display expanded polygons on both the hypertension and obesity vertices, but diverge on the insomnia axis: Cluster 1 (insomniac) extends to the maximum along the insomnia vertex, while Cluster 2 (non-insomniac) collapses inward on that same vertex. Clusters 4 and 6 (hypertensive, overweight) present polygons that are elongated along the hypertension and overweight axes but contracted on the obesity vertices, again splitting on insomnia. Clusters 3 and 5 (normotensive) present the most compact polygons on the hypertension axis, with their profiles concentrated along the overweight and insomnia dimensions.

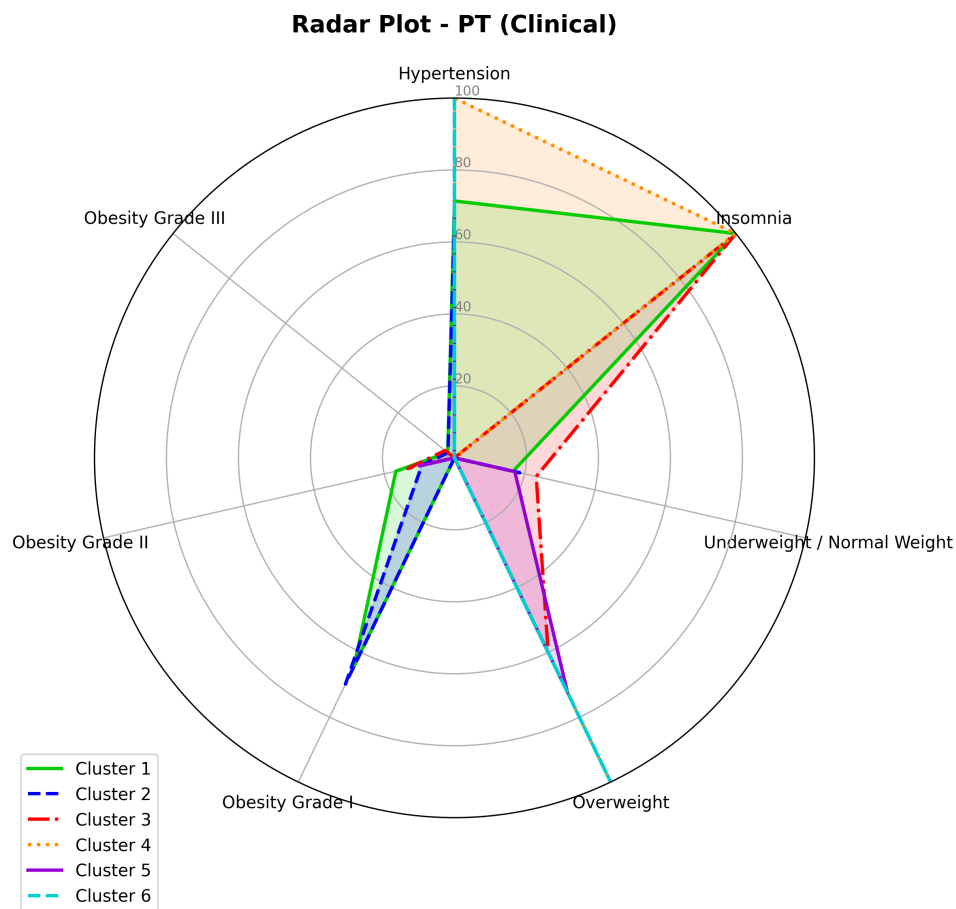


Figure 4: Radar plot of the six clusters identified in the Portuguese cohort (clinical approach). Each polygon represents a cluster's multivariate profile across the discriminating variables.

4.1.2 Spanish dataset

4.1.2.1 Baseline characteristics

The Spanish cohort comprised 355 patients, of whom 287 (80.8%) had a confirmed diagnosis of OSA and 68 (19.2%) were classified as healthy. As described in Table 4, the cohort was predominantly male in both groups (72.1% healthy; 76.0% OSA; $p = 0.608$).

In contrast to the Portuguese cohort, age was a strong discriminator in the Spanish dataset ($p < 0.001$). The proportion of patients in the youngest age stratum (20–44) was markedly higher in the healthy group (27/68, 39.7%) compared to the OSA group (39/287, 13.6%), while middle aged (45–65) and older (66–90) patients were overrepresented among those with OSA. BMI distribution also differed significantly ($p < 0.001$): normal weight patients accounted for 36.8% of the healthy group but only 16.0% of the OSA group, whereas Obesity Grades II and III were substantially more prevalent in the OSA group (13.9% and 8.0%, respectively, vs. 4.4% and 0.0% in the healthy group).

Arterial hypertension was significantly more prevalent in the OSA group (124/287, 43.2% vs. 15/68, 22.1%; $p = 0.002$), as was diabetes (59/287, 20.6% vs. 5/68, 7.4%; $p = 0.018$). These two comorbidities, unlike in the Portuguese cohort where only hypertension reached significance, both emerged as significant markers in the Spanish dataset. Notably, excessive daytime sleepiness was highly prevalent in both groups (44.1% healthy; 48.4% OSA; $p = 0.613$), a pattern distinct from the Portuguese cohort where EDS was comparatively rare ($<18\%$ in both groups). Insomnia, in contrast, showed no significant difference between groups in the Spanish cohort ($p > 0.999$), a stark contrast to the Portuguese dataset where it was one of the strongest discriminators.

Table 4: Descriptive characteristics of the Spanish study cohort, n (%). P-values obtained using the chi square test unless otherwise specified.

Variable	Healthy (n=68)	OSA (n=287)	p value
Male gender	49 (72.1)	218 (76.0)	0.608
Age			<0.001
20–44	27 (39.7)	39 (13.6)	
45–65	29 (42.6)	171 (59.6)	
66–90	12 (17.6)	77 (26.8)	
BMI (kg/m²)			<0.001
Underweight / Normal	25 (36.8)	46 (16.0)	
Overweight	28 (41.2)	111 (38.7)	
Obesity Grade I	12 (17.6)	67 (23.3)	
Obesity Grade II	3 (4.4)	40 (13.9)	
Obesity Grade III	0 (0)	23 (8.0)	
Alcohol	9 (13.2)	61 (21.3)	0.185
Tobacco	30 (44.1)	165 (57.5)	0.063
Bariatric surgery	0 (0)	8 (2.8)	0.362 [#]
Depression	0 (0)	1 (0.3)	>0.999 [#]
Hypertension	15 (22.1)	124 (43.2)	0.002
Cerebrovascular	3 (4.4)	15 (5.2)	>0.999 [#]
Asthma	11 (16.2)	27 (9.4)	0.160
COPD	3 (4.4)	17 (5.9)	0.776 [#]
Diabetes	5 (7.4)	59 (20.6)	0.018
Dyslipidemia	30 (44.1)	146 (50.9)	0.386
Thyroid disease	11 (16.2)	34 (11.8)	0.446
Dementia	3 (4.4)	2 (0.7)	0.050 [#]
Cancer	5 (7.4)	30 (10.5)	0.586
GE reflux	16 (23.5)	43 (15.0)	0.128
Glaucoma	1 (1.5)	6 (2.1)	>0.999 [#]
Epworth (ESS)			0.613
Normal	38 (55.9)	148 (51.6)	
EDS	30 (44.1)	139 (48.4)	
Gasping	18 (26.5)	67 (23.3)	0.700
Snoring	56 (82.4)	250 (87.1)	0.408
Witnessed apneas	35 (51.5)	161 (56.1)	0.579
Insomnia	11 (16.2)	48 (16.7)	>0.999

[#] Fisher exact test. Abbreviations: BMI, body mass index; COPD, chronic obstructive pulmonary disease; ESS, Epworth Sleepiness Scale; EDS, excessive daytime sleepiness; GE, gastroesophageal. Significant p values (p < 0.05) are shown in bold.

4.1.2.2 Cluster characterization

The k-modes algorithm identified four clusters in the Spanish cohort, as detailed in Table 5. In contrast to the six cluster solution found in the Portuguese dataset, the four cluster solution in the Spanish cohort reflects the stronger role of age as a partitioning variable and the absence of insomnia as a discriminator.

Cluster 1 ($n = 117$, 33.0%) was the “elderly” cluster, with 76.1% of patients aged 66–90. This cluster exhibited a high prevalence of hypertension (84/117, 71.8%) and the highest prevalence of diabetes (33/117, 28.2%) among all groups and the modal BMI was overweight (56.4%). Cluster 2 ($n = 124$, 34.9%) was entirely composed of middle aged patients (45–65; 100%), with no hypertension (0%) and a lower diabetes prevalence (12.1%). This cluster showed a broad BMI distribution, with overweight (40.3%) and Obesity Grade I (25.0%) as the dominant categories.

Cluster 3 ($n = 55$, 15.5%) was the smallest cluster and represented a “middle aged obese hypertensive” phenotype: 98.2% of patients were aged 45–65, all had hypertension (100%), and the BMI distribution was heavily skewed toward higher obesity grades (Obesity Grade II: 40.0%; Grade III: 16.4%). Cluster 4 ($n = 59$, 16.6%) was exclusively composed of young patients aged 20–44 (100%), with no hypertension and minimal diabetes (3.4%), presenting the highest proportion of normal weight individuals (28.8%) among all clusters.

Table 5: Clinical characteristics of the Spanish cohort by the defined clusters, n (%).

Variable	C1 ($n=117$)	C2 ($n=124$)	C3 ($n=55$)	C4 ($n=59$)
Hypertension	84 (71.8)	0 (0)	55 (100)	0 (0)
Diabetes	33 (28.2)	15 (12.1)	14 (25.5)	2 (3.4)
Age				
20–44	6 (5.1)	0 (0)	1 (1.8)	59 (100)
45–65	22 (18.8)	124 (100)	54 (98.2)	0 (0)
66–90	89 (76.1)	0 (0)	0 (0)	0 (0)
BMI				
UW / Normal	22 (18.8)	26 (21.0)	6 (10.9)	17 (28.8)
Overweight	66 (56.4)	50 (40.3)	0 (0)	23 (39.0)
Obesity I	18 (15.4)	31 (25.0)	18 (32.7)	12 (20.3)
Obesity II	5 (4.3)	11 (8.9)	22 (40.0)	5 (8.5)
Obesity III	6 (5.1)	6 (4.8)	9 (16.4)	2 (3.4)

Abbreviations: BMI, body mass index; UW, underweight; C, cluster.

4.1.2.3 Cluster profiles and phenotype definition

Table 6 summarizes the modal clinical profiles. A particularly noteworthy finding was that all four clusters exhibited a modal AHI severity of severe, with mean AHI values ranging from 34.3 (Cluster 4) to 56.6 (Cluster 3). This uniformly severe distribution contrasts with the Portuguese cohort, where modal severities ranged from mild to severe.

Cluster 3 (middle aged, obese, universally hypertensive) had the highest mean AHI of any cluster across both cohorts (56.6 ± 32.8), identifying it as the highest risk group. Cluster 4 (young, normotensive) had the lowest mean AHI (34.3 ± 32.7), though the high standard deviation indicates substantial within cluster heterogeneity.

Table 6: Modes of the clustered clinical presentations of OSA – Spanish cohort (clinical approach).

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
<i>Demographics</i>				
Age	66–90	45–65	45–65	20–44
BMI	Overweight	Overweight	Obesity II	Overweight
<i>Comorbidities</i>				
Hypertension	Yes	No	Yes	No
Diabetes	No	No	No	No
<i>AHI</i>				
Mean (SD)	47.7 (27.0)	41.2 (28.2)	56.6 (32.8)	34.3 (32.7)
Median (IQR)	47.6	35.0	52.7	22.6
Severity	Severe	Severe	Severe	Severe

Modes represent the most frequent category per variable within each cluster. AHI was not used as a clustering variable.

4.1.2.4 Visualization of cluster profiles

The heatmap in Fig. 5 illustrates the distinct age and comorbidity profiles of the four clusters. The age variable partitions the clusters most clearly: Cluster 1 concentrates in the 66–90 stratum, Clusters 2 and 3 in the 45–65 stratum, and Cluster 4 exclusively in the 20–44 stratum. The hypertension and diabetes axes further differentiate the hypertensive clusters (C1 and C3) from the normotensive clusters (C2 and C4).

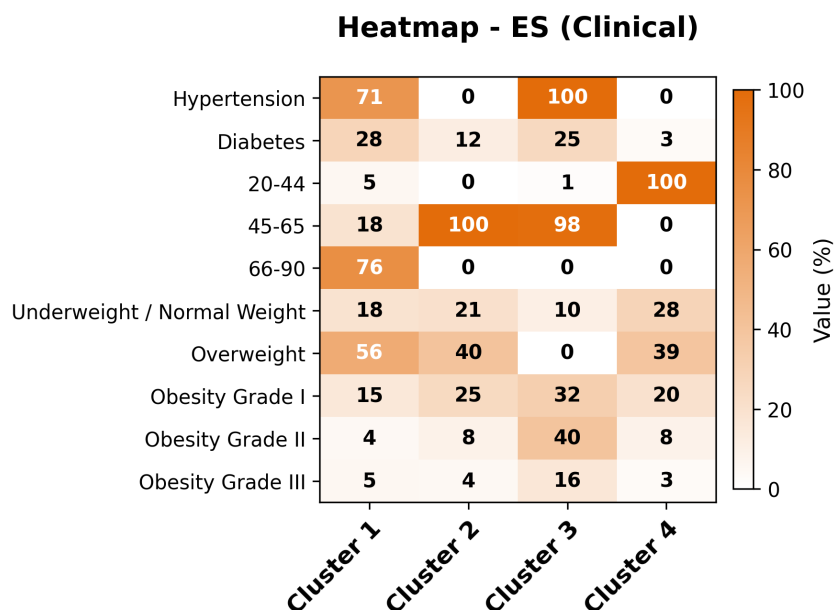


Figure 5: Heatmap of categorical variable distributions across the four clusters identified in the Spanish cohort (clinical approach).

The radar plot in Fig. 6 maps each cluster's profile across the discriminating axes: hypertension, diabetes, age strata (20–44, 45–65, 66–90), and BMI categories. In contrast to the six-polygon Portuguese solution, the four Spanish clusters produce more differentiated polygons, primarily driven by the age and hypertension axes. Cluster 1 (elderly hypertensive) and Cluster 3 (middle-aged obese hypertensive) both exhibit expanded cardiovascular footprints, but Cluster 3 extends further along the metabolic axis (BMI), reflecting its concentration of Obesity Grades II and III. Cluster 4 (young normotensive) presents the most compact polygon, consistent with its minimal comorbidity burden and lowest mean AHI (34.3). Cluster 2 (middle-aged normotensive) occupies an intermediate position, with a moderately expanded polygon but no cardiovascular extension.

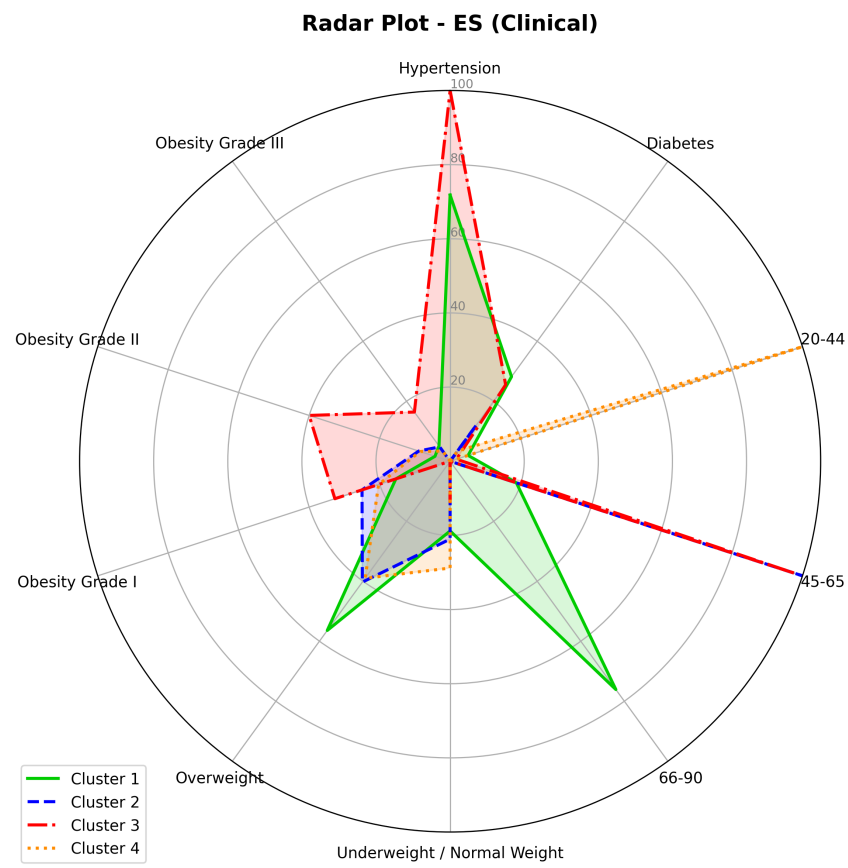


Figure 6: Radar plot of the four clusters identified in the Spanish cohort (clinical approach). Each polygon represents a cluster's multivariate profile across the discriminating variables.

4.2 Scientific approach

Following the data-driven variable selection strategy described in Section 3.3.2, the k-modes algorithm was applied independently to both cohorts using the full set of 23 pre-triage clinical variables. The results for each cohort are presented below.

4.2.1 Portuguese dataset

4.2.1.1 Baseline characteristics

The underlying cohort is identical to that described in Table 1 ($n = 188$); the differences between the clinical and scientific approaches reside exclusively in variable selection and preprocessing.

4.2.1.2 Cluster characterization

The k-modes algorithm, applied with the full variable set under the scientific approach, identified a two cluster solution in the Portuguese cohort, in contrast to the six cluster solution obtained under the clinical approach. Table 7 presents the detailed clinical characteristics.

Cluster 1 ($n = 113$, 60.1%) was predominantly female (63.7%) and older, with 54.0% of patients aged 66–90. This cluster carried a high comorbidity burden: hypertension (72.6%), dyslipidemia (70.8%), diabetes (29.2%), cerebrovascular disease (19.5%), and depression (23.9%). Insomnia was highly prevalent (63.7%). Alcohol consumption was lower (34.5%) and tobacco use was low (12.4%).

Cluster 2 ($n = 75$, 39.9%) was predominantly male (81.3%) and younger, with 73.3% of patients aged 45–65 and 25.3% aged 20–44. This cluster presented a markedly lower comorbidity burden: hypertension (28.0%), dyslipidemia (30.7%), diabetes (14.7%), and cerebrovascular disease (1.3%). Alcohol consumption was higher (72.0%) and insomnia was less prevalent (22.7%). Both clusters showed similar Epworth distributions (approximately 83% normal in both) and near-universal snoring ($\geq 97\%$).

Table 7: Clinical characteristics of the Portuguese cohort by the defined clusters (scientific approach), n (%).

Variable	C1 (n=113)	C2 (n=75)
Male gender	41 (36.3)	61 (81.3)
Age		
20–44	8 (7.1)	19 (25.3)
45–65	44 (38.9)	55 (73.3)
66–90	61 (54.0)	1 (1.3)
BMI		
Normal	16 (14.2)	11 (14.7)
Overweight	51 (45.1)	32 (42.7)
Obesity I	34 (30.1)	23 (30.7)
Obesity II	10 (8.8)	8 (10.7)
Obesity III	2 (1.8)	1 (1.3)
Alcohol	39 (34.5)	54 (72.0)
Tobacco	14 (12.4)	22 (29.3)
Bariatric surgery	3 (2.7)	3 (4.0)
Depression	27 (23.9)	9 (12.0)
Hypertension	82 (72.6)	21 (28.0)
Cerebrovascular	22 (19.5)	1 (1.3)
Asthma	15 (13.3)	9 (12.0)
COPD	11 (9.7)	6 (8.0)
Diabetes	33 (29.2)	11 (14.7)
Dyslipidemia	80 (70.8)	23 (30.7)
Thyroid disease	21 (18.6)	5 (6.7)
Dementia	1 (0.9)	0 (0)
Cancer	19 (16.8)	7 (9.3)
GE reflux	2 (1.8)	3 (4.0)
Glaucoma	1 (0.9)	2 (2.7)
Epworth (ESS)		
Normal	95 (84.1)	62 (82.7)
EDS	18 (15.9)	13 (17.3)
Gasping	6 (5.3)	8 (10.7)
Snoring	113 (100.0)	73 (97.3)
Witnessed apneas	97 (85.8)	69 (92.0)
Insomnia	72 (63.7)	17 (22.7)

Abbreviations: BMI, body mass index; COPD, chronic obstructive pulmonary disease; ESS, Epworth Sleepiness Scale; EDS, excessive daytime sleepiness; GE, gastroesophageal; C, cluster.

4.2.1.3 Cluster profiles and phenotype definition

Table 8 summarizes the modal clinical profiles. The two clusters clearly separated along axes of age, gender, and comorbidity. Cluster 1 constitutes an “older, highly comorbid, predominantly female” phenotype: elderly, overweight, with high rates of hypertension, dyslipidemia, and insomnia. In contrast, Cluster 2 represents a “younger, lower comorbidity, predominantly male” phenotype: middle-aged, overweight, non-hypertensive, non-dyslipidemic, but with higher alcohol consumption.

From an AHI perspective, Cluster 1 presented a higher mean AHI (25.5 ± 19.9) compared to Cluster 2 (22.1 ± 20.3), with a modal severity of severe (37.2%) versus mild (32.0%) in Cluster 2. This gradient aligns with the accumulation of cardiometabolic risk factors in the older phenotype.

Table 8: Modes of the clustered clinical presentations of OSA – Portuguese cohort (scientific approach).

	Cluster 1	Cluster 2
<i>Demographics</i>		
Male gender	No	Yes
Age	66–90	45–65
<i>Behavior</i>		
BMI	Overweight	Overweight
Alcohol	No	Yes
Tobacco	No	No
<i>Comorbidities</i>		
Hypertension	Yes	No
Diabetes	No	No
Dyslipidemia	Yes	No
<i>Symptoms</i>		
Epworth (ESS)	Normal	Normal
Witnessed apneas	Yes	Yes
Snoring	Yes	Yes
Insomnia	Yes	No
<i>AHI</i>		
Mean (SD)	25.5 (19.9)	22.1 (20.3)
Median (IQR)	21.7 (9.2–37.9)	15.1 (7.1–28.0)
Severity (mode)	Severe	Mild

Modes represent the most frequent category per variable within each cluster. AHI was not used as a clustering variable.

4.2.1.4 Visualization of cluster profiles

Fig. 7 presents the heatmap of the two-cluster partition in the Portuguese cohort. The color intensity map reveals that the two clusters diverge most sharply on age (Cluster 1 concentrated in

the 66–90 stratum; Cluster 2 in the 20–44 and 45–65 strata), comorbidity burden (hypertension, dyslipidemia, and diabetes elevated in Cluster 1), and behavioral variables (alcohol and tobacco consumption elevated in Cluster 2).

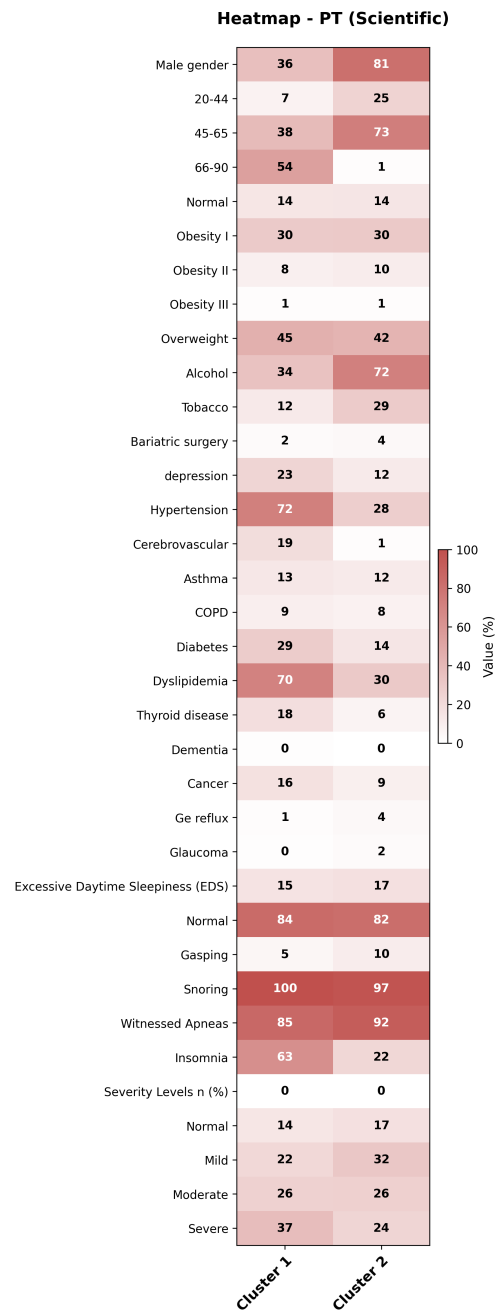


Figure 7: Heatmap of categorical variable distributions across the two clusters identified in the Portuguese cohort (scientific approach).

4.2.2 Spanish dataset

4.2.2.1 Baseline characteristics

The underlying cohort for the scientific approach is identical to that described in Table 4: the same 355 patients with the same demographic and clinical profiles. The distinction between the two approaches lies exclusively in the number and nature of the variables selected for clustering.

4.2.2.2 Cluster characterization

The k-modes algorithm, applied with the full variable set, identified a two-cluster solution in the Spanish cohort. Table 9 presents the detailed clinical characteristics.

Cluster 1 (n = 189, 53.2%) was predominantly male (75.7%) and younger, with 54.0% of patients aged 45–65 and 33.9% aged 20–44. This cluster presented a markedly lower comorbidity burden: hypertension (14.8%), dyslipidemia (29.1%), and diabetes (9.5%). Excessive daytime sleepiness (EDS) and witnessed apneas were highly prevalent (67.7% and 69.3%, respectively). Alcohol consumption was lower (14.8%) and tobacco use was moderate (33.3%).

Cluster 2 (n = 166, 46.8%) was also predominantly male (74.7%) with 59.0% of patients aged 45–65 and 39.8% aged 66–90. This cluster carried a high comorbidity burden: hypertension (66.9%), dyslipidemia (72.9%), and diabetes (27.7%). Excessive daytime sleepiness (EDS) and witnessed apneas were less prevalent (24.7% and 39.2%, respectively). Alcohol consumption was higher (25.3%) and tobacco use was high (79.5%). Both clusters showed high prevalence of snoring (90.5% in C1; 81.3% in C2) and similar insomnia distributions (approximately 16% in both).

Table 9: Clinical characteristics of the Spanish cohort by the defined clusters (scientific approach), n (%).

Variable	C1 (n=189)	C2 (n=166)
Male gender	143 (75.7)	124 (74.7)
Age		
20–44	64 (33.9)	2 (1.2)
45–65	102 (54.0)	98 (59.0)
66–90	23 (12.2)	66 (39.8)
BMI		
UW / Normal	42 (22.2)	29 (17.5)
Overweight	82 (43.4)	57 (34.3)
Obesity I	36 (19.0)	43 (25.9)
Obesity II	17 (9.0)	26 (15.7)
Obesity III	12 (6.3)	11 (6.6)
Alcohol	28 (14.8)	42 (25.3)
Tobacco	63 (33.3)	132 (79.5)
Bariatric surgery	2 (1.1)	6 (3.6)
Depression	0 (0)	1 (0.6)
Hypertension	28 (14.8)	111 (66.9)
Cerebrovascular	6 (3.2)	12 (7.2)
Asthma	23 (12.2)	15 (9.0)
COPD	4 (2.1)	16 (9.6)
Diabetes	18 (9.5)	46 (27.7)
Dyslipidemia	55 (29.1)	121 (72.9)
Thyroid disease	23 (12.2)	22 (13.3)
Dementia	1 (0.5)	4 (2.4)
Cancer	11 (5.8)	24 (14.5)
GE reflux	23 (12.2)	36 (21.7)
Glaucoma	4 (2.1)	3 (1.8)
Epworth (ESS)		
Normal	61 (32.3)	125 (75.3)
EDS	128 (67.7)	41 (24.7)
Gasping	53 (28.0)	32 (19.3)
Snoring	171 (90.5)	135 (81.3)
Witnessed apneas	131 (69.3)	65 (39.2)
Insomnia	31 (16.4)	28 (16.9)

Abbreviations: BMI, body mass index; COPD, chronic obstructive pulmonary disease; ESS, Epworth Sleepiness Scale; EDS, excessive day-time sleepiness; GE, gastroesophageal; UW, underweight; C, cluster.

4.2.2.3 Cluster profiles and phenotype definition

Table 10 presents the modal clinical profiles of the two clusters. Building upon the detailed characterization from the previous section, the two groups can be synthesized into distinct phenotypes:

Cluster 1 constitutes a “symptomatic, low comorbidity” phenotype, whereas Cluster 2 represents an “asymptomatic, high comorbidity” phenotype.

Despite these marked differences in clinical presentation, the two clusters exhibited virtually identical AHI distributions: mean AHI of 44.4 ± 31.3 (Cluster 1) versus 44.8 ± 28.5 (Cluster 2), with median values of 39.4 and 39.3, respectively. The modal severity was severe in both groups (61.4% and 59.6%, respectively). This dissociation between clinical phenotype and AHI severity reinforces the argument that AHI alone fails to capture the heterogeneity of OSA, as two populations with nearly identical severity indices present fundamentally different clinical profiles and, potentially, different treatment needs.

Table 10: Modes of the clustered clinical presentations of OSA – Spanish cohort (scientific approach).

	Cluster 1	Cluster 2
<i>Demographics</i>		
Male gender	Yes	Yes
Age	45–65	45–65
<i>Behavior</i>		
BMI	Overweight	Overweight
Alcohol	No	No
Tobacco	No	Yes
<i>Comorbidities</i>		
Hypertension	No	Yes
Diabetes	No	No
Dyslipidemia	No	Yes
<i>Symptoms</i>		
Epworth (ESS)	EDS	Normal
Witnessed apneas	Yes	No
Snoring	Yes	Yes
Insomnia	No	No
<i>AHI</i>		
Mean (SD)	44.4 (31.3)	44.8 (28.5)
Median (IQR)	39.4 (17.7–65.1)	39.3 (20.6–65.6)
Severity (mode)	Severe	Severe

Modes represent the most frequent category per variable within each cluster. AHI was not used as a clustering variable.

4.2.2.4 Visualization of cluster profiles

The heatmap in Fig. 8 provides a visual summary of the two-cluster partition in the Spanish cohort. The color intensity map highlights the sharp divergence between the two clusters on the behavioral axis (tobacco: 33.3% vs. 79.5%), the cardiovascular axis (hypertension: 14.8% vs. 66.9%;

dyslipidemia: 29.1% vs. 72.9%), and the symptomatic axis (EDS: 67.7% vs. 24.7%; witnessed apneas: 69.3% vs. 39.2%).

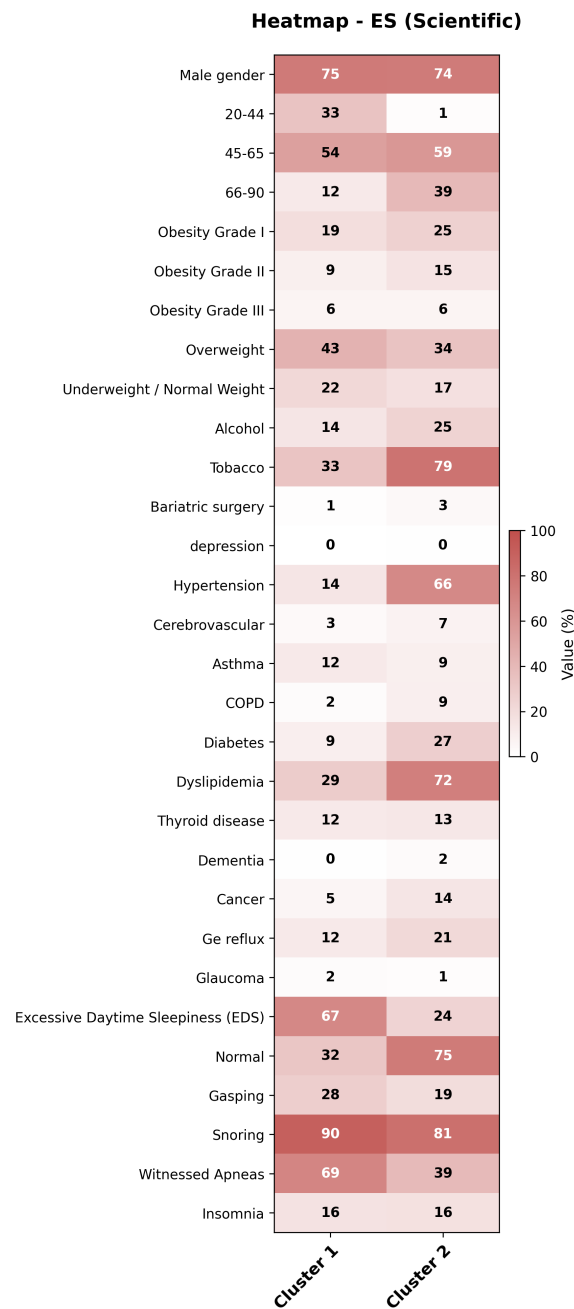


Figure 8: Heatmap of categorical variable distributions across the two clusters identified in the Spanish cohort (scientific approach).

4.2.3 Comparative radar plot analysis

To complement the heatmap representations and facilitate cross-cohort comparison of the scientific approach clusters, radar plots were generated for both the Portuguese and Spanish two-cluster solutions. Each radar plot maps the cluster profiles as polygons whose vertices correspond to the principal axes of clinical variation identified by the data-driven approach: age strata (20–44, 66–90), behavioral variables (tobacco), cardiovascular comorbidities (hypertension, diabetes, dyslipidemia), and symptom dimensions (excessive daytime sleepiness, witnessed apneas).

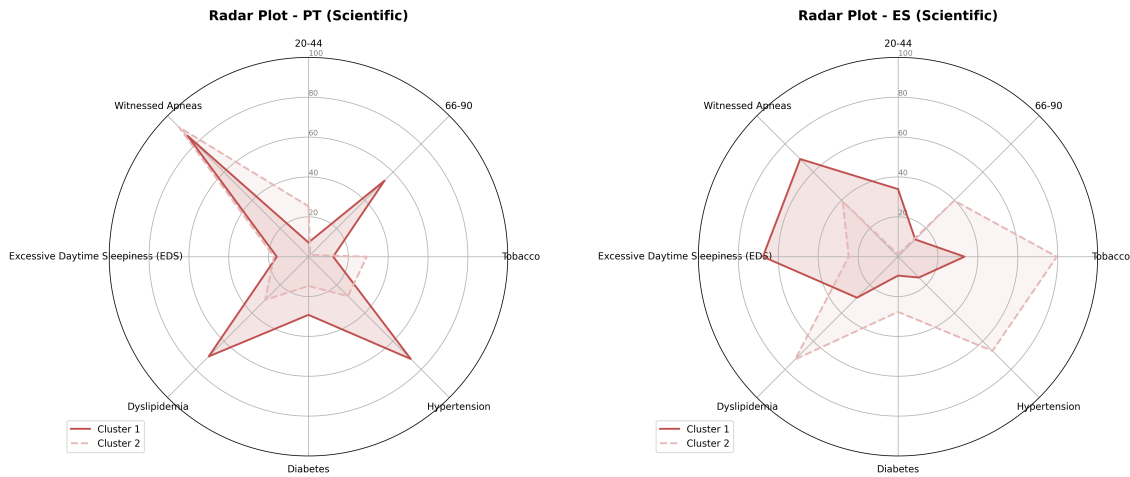


Figure 9: Radar plots generated for both the Portuguese (left) and Spanish (right) two-cluster solutions. Each polygon represents a cluster’s multivariate profile across all 23 clinical variables.

A side-by-side comparison of the radar plots for the Portuguese and Spanish cohorts (Fig. 9) reveals a fundamental structural convergence. In the Portuguese cohort, Cluster 1 displays a polygon globally expanded along the cardiovascular and age 66–90 axes, consistent with higher AHI severity, while Cluster 2 presents a more compact polygon. In the Spanish cohort, despite identical AHI distributions (mean ≈ 44), the two polygons exhibit an immediate visual “mirror image” morphotypic inversion: Cluster 1 presents an expanded symptomatic axis but a contracted cardiovascular axis, while Cluster 2 exhibits the exact opposite.

Despite these local nuances, particularly the stronger role of symptomatic axes in the Spanish data, the data-driven scientific approach identified an equivalent number of clusters ($k = 2$) in both cohorts. Critically, the two clusters obtained in each cohort encode the same underlying phenotypic axis: one cluster captures an “older, highly comorbid” profile (Portuguese C1, Spanish C2), characterized by elevated hypertension, dyslipidemia, and diabetes prevalence, while the other captures a “younger, lower comorbidity” profile (Portuguese C2, Spanish C1), characterized by lower cardiovascular burden. This convergence across two independent cohorts from different Iberian populations constitutes evidence of reproducibility of the data-driven partition, suggesting a robust, population-level axis of OSA heterogeneity.

Chapter 5

Discussion

This chapter provides a critical interpretation of the clustering results presented in Chapter 4, contextualizing the identified phenotypes within the existing literature and discussing their clinical implications. The discussion is organized around four axes: (i) the cohort-level comparison and the referral bias that fundamentally shapes the two datasets (Section 5.1); (ii) an exhaustive interpretation of the phenotypic structures identified under the clinical and scientific approaches, including the patterns revealed by heatmaps and radar plots, and their comparison with the broader phenotyping literature (Section 5.2); (iii) the confrontation between the two analytical strategies, including methodological and visualization considerations (Section 5.3); and (iv) the limitations of the study (Section 5.4).

5.1 Cohort comparison and selection bias

5.1.1 Inter-cohort discrepancies

The Portuguese cohort was smaller and exhibited a more balanced sex distribution, whereas the Spanish cohort was predominantly male and carried a markedly higher overall AHI burden. This male predominance in the Spanish cohort mirrors the demographic profile reported in the majority of large-scale phenotyping studies, including Bailly et al. (2016) [16], who reported 72.8% male prevalence across 18,263 French OSA patients, and Keenan et al. (2018) [17], who observed comparable ratios across five international sleep centers. In contrast, the more balanced sex distribution in the Portuguese cohort is consistent with the demographic profile observed by Ferreira-Santos and Rodrigues (2023) [26], who reported a substantial female representation in their Portuguese cohort from the same regional healthcare system. This Portuguese sex balance may partly reflect the diagnostic pathway: in-lab PSG referrals in the Portuguese center tend to include patients with more complex clinical presentations, among whom female patients with COMISA (co-morbid insomnia and sleep apnea) and atypical symptom profiles may be overrepresented relative to community prevalence. The relevance of sex-specific phenotyping has been underscored by Fontanilles-Arbones et al. (2024) [3], who applied divisive hierarchical clustering to a female-specific cohort of 1,886 patients and identified four distinct phenotypes, demonstrating that sex-stratified analyses can reveal phenotypic structures that are obscured in mixed-sex cohorts.

Notably, the Portuguese cohort showed no statistically significant age difference between healthy and OSA groups, whereas age was one of the strongest discriminators in the Spanish dataset, with the OSA group being significantly older. Both cohorts converged on the significance of BMI and hypertension, confirming these variables as robust cross-population markers of OSA risk. This convergence is consistent with the findings of Bailly et al. (2016) [16], who

identified hypertension as the most prevalent comorbidity across all six of their clusters (overall 46.4%), and with the multi-center analysis by Testelmans et al. (2022) [37], who reported hypertension and obesity as the two most prevalent objectively assessed comorbidities in their Belgian cohort (75% and 63%, respectively). Pevernagie (2021) [21] further reinforced the centrality of these cardiometabolic markers in the conceptual framework of treatable traits in sleep-disordered breathing.

A critical divergence was observed in the role of diabetes: statistically significant in the Spanish cohort but not in the Portuguese. This asymmetry is particularly noteworthy because diabetes has emerged as a consistent discriminator in several large European phenotyping studies. Bailly et al. (2016) [16] reported diabetes prevalences ranging from 0.3% (their young snorer cluster) to 38.8% (their multi-disease elderly obese cluster), while Testelmans et al. (2022) [37] identified a dedicated “diabetes” cluster (Cluster C4, 9.2% of patients) characterized by higher prevalence and degree of obesity and diabetes alongside elevated dyslipidemia. The absence of significance in the Portuguese cohort may reflect genuine epidemiological differences between the Iberian populations, or, more plausibly, differences in clinical documentation depth and the smaller sample size that limits statistical power for detecting moderate effect sizes.

Conversely, insomnia emerged as a significant differentiator exclusively in the Portuguese cohort, a finding with important clinical implications. This may reflect both screening practices and the higher sensitivity of Portuguese clinicians to COMISA presentations, as highlighted by Miranda et al. (2025) [59]. Insomnia, despite its high prevalence in OSA populations, remains a relatively underutilized variable in phenotyping studies. Among the major clustering studies reviewed, only Ferreira-Santos and Rodrigues (2021) [6] explicitly incorporated insomnia-related variables, and their phenotypes were partly driven by non-repairing sleep and morning headaches. The prominence of insomnia as a stratifying variable in our Portuguese cohort adds novel evidence supporting the relevance of COMISA-related symptoms in OSA phenotyping and echoes the growing recognition that sleep quality dimensions may capture clinically meaningful heterogeneity that purely respiratory or cardiometabolic variables overlook [31].

5.1.2 Referral bias and diagnostic pathway asymmetry

The most striking inter-cohort difference concerns the AHI severity distribution. In the clinical approach, all four Spanish clusters exhibited a modal AHI severity of severe (mean range: 34.3–56.6), whereas the Portuguese clusters spanned the full severity spectrum from mild (mean AHI 14.9) to severe (mean AHI 33.3). In the scientific approach, the same pattern persisted: the two Spanish clusters had virtually identical mean AHI values (≈ 44), while the two Portuguese clusters differed meaningfully (25.5 vs. 22.1).

This discrepancy is not a clinical artefact; it is a direct consequence of the fundamentally different diagnostic pathways operating in each country. As documented by Miranda et al. (2025) [59], the Spanish center (Hospital Universitario Río Hortega, Valladolid) employs Level I polysomnography (in-lab PSG) as the first-line diagnostic examination, offering it broadly to any patient presenting with sleep-related complaints. This practice is facilitated by greater resource availability,

resulting in a referral population that is representative of the full severity spectrum of OSA in the community.

By contrast, the Portuguese center Unidade Local de Saúde de São João operates under substantial resource constraints that impose a tiered diagnostic pathway. General screening is performed via HSAT studies, which captures the majority of patients. In-lab PSG is reserved for a highly selected subset of patients: those in whom HSAT was inconclusive, those requiring absolute diagnostic certainty (e.g., pre-surgical evaluation), or those with severe neurological or psychiatric comorbidities (e.g., dementia, severe anxiety) that preclude reliable home monitoring. As a consequence, the patients who populate our Portuguese dataset represent a pre-selected, clinically complex subgroup rather than a random sample of the population referred for sleep evaluation. This selection mechanism biases the Portuguese dataset toward lower mean AHI values, as the most severe, straightforward cases are often diagnosed and managed through HSAT without ever reaching the in-lab PSG setting.

This phenomenon, known as referral or spectrum bias [60], has profound implications for the interpretation of our clustering results. The broader severity spectrum in the Portuguese cohort (mild to severe across clusters) likely reflects the inclusion of patients whose HSAT results were equivocal, many of whom may have borderline or mild disease. The uniformly severe Spanish profile, conversely, reflects a referral population in which in-lab PSG is the default pathway, capturing the full range of community severity but with a higher overall baseline. This asymmetry is not unique to our study: Silveira et al. (2022) [7], in their longitudinal follow-up of OSA phenotypes from a Spanish center, reported a similarly high overall AHI burden across all three of their clusters, with mean AHI values ranging from 32 to 55, consistent with the population-level severity profile observed in our Spanish cohort. Miranda et al. (2025) [59] explicitly noted this asymmetry, observing that “the overall higher AHI values, and consequently greater OSA severity, in the Spanish population may be attributed to differences in diagnostic procedures,” with in-lab PSG serving as the first-line exam in Spain versus HSAT in Portugal. Our clustering results are entirely consistent with this interpretation: the phenotypic structure of each cohort is shaped not only by the underlying clinical heterogeneity but also by the diagnostic pathway through which patients were selected

5.2 Phenotypic interpretation and comparison with the literature

5.2.1 Clinical approach: cohort-specific phenotypic drivers

The application of the clinical approach yielded distinct phenotypic structures for each cohort, highlighting that although both Iberian populations share fundamental cardiometabolic risk factors, their clinical partitions were driven by different primary axes.

The Spanish cohort: age and cardiometabolic drivers. In the Spanish cohort, the algorithm identified an optimal four-cluster solution primarily organized along age and cardiometabolic severity. This structural convergence provides cross-algorithmic validation of the four-phenotype

model previously identified in Spanish OSA populations by Romero-Peralta et al. (2022) [28]. Furthermore, the age-driven nature of these Spanish clusters aligns closely with the international analysis by Keenan et al. (2018) [17], who found that age and BMI were among the strongest discriminators across multiple international sleep centers.

The Portuguese cohort: the role of insomnia and COMISA. However, while Keenan et al. (2018) found their clusters to be largely replicable across centers, our Portuguese cohort yielded a distinct, more granular six-cluster solution. A key distinguishing feature in Portugal was the independent contribution of insomnia to AHI severity. Within identical cardiometabolic strata, insomniac clusters exhibited substantially higher AHI burdens than non-insomniac clusters (e.g., C4 vs. C6, and C1 vs. C2). This pattern aligns with the COMISA literature, echoing Wei et al. (2023) [31] in demonstrating that sleep quality dimensions capture clinically meaningful heterogeneity. Contextualizing this within national data, our six-cluster solution serves as a finer-grained decomposition of the three-cluster partition reported by Ferreira-Santos and Rodrigues (2023) [26], emphasizing that the inclusion of insomnia variables provides essential stratification resolution in clinically complex referral populations.

5.2.2 Scientific approach: cross-cohort convergence on a comorbidity–age axis

Despite local nuances, particularly the stronger role of symptomatic axes (EDS, witnessed apneas) in the Spanish data and the stronger role of insomnia in the Portuguese data, the underlying comorbidity–age axis was conserved across both populations. This convergence, which would require careful row-by-row comparison in a heatmap format, was immediately apparent in the radar representation through the similar polygon shapes and area distributions, underscoring the value of multivariate geometric visualization for cross-cohort phenotypic comparison. This cross-cohort convergence is of particular importance because it was achieved through an entirely data-driven approach, without any a priori clinical hypothesis guiding variable selection. The identification of the same fundamental axis of OSA heterogeneity in two independent Iberian populations strengthens the argument that the comorbidity–age axis represents a robust, population-level dimension of clinical variation in OSA. This finding is consistent with several independent lines of evidence in the phenotyping literature:

- Bailly et al. (2016) [16], in their six-cluster analysis of 18,263 French patients, identified two fundamental axes of variation: (i) a comorbidity axis separating clusters with few or no comorbidities (Clusters 1, 2, 4) from multi-disease clusters (Clusters 3, 6), and (ii) a symptom axis separating symptomatic from asymptomatic presentations. The former axis is directly analogous to the comorbidity–age axis identified by our scientific approach.
- Testelmans et al. (2022) [37], using self-organizing maps to cluster 710 Belgian OSA patients based on 10 objectively identified comorbidities, identified nine clusters that could be broadly grouped into high-comorbidity (e.g., their “diabetes” and “obesity–hypertension” clusters) and low-comorbidity (their “least comorbidities” cluster) phenotypes, reinforcing the primacy of the comorbidity dimension in OSA phenotypic architecture.

- Vavougiou et al. (2016) [35], applying categorical principal component analysis combined with cluster analysis to Greek OSA patients, similarly identified comorbidity burden as the primary axis of phenotypic variation, with clusters differentiated by the presence and combination of cardiovascular, metabolic, and inflammatory comorbidities.
- Ntenta et al. (2022) [36], extending the methodology of Vavougiou et al. to a primary healthcare setting in Northern Greece, identified OSA phenotypes characterized by distinct comorbidity profiles, demonstrating that the comorbidity–age axis is detectable even in primary care populations where OSA severity is typically milder.

5.2.2.1 The dissociation between clinical phenotype and AHI severity

Perhaps the most clinically significant finding of the scientific approach was the dissociation between clinical phenotype and AHI severity observed in the Spanish cohort. Despite exhibiting identical AHI distributions, the two Spanish clusters presented markedly distinct clinical profiles: one symptomatic with low cardiometabolic burden, the other asymptomatic with high cardiometabolic burden. This finding has profound implications for the current classification of OSA, which remains overwhelmingly dependent on AHI as the primary metric of disease severity.

The dissociation we observed is consistent with mounting evidence that AHI fails to capture the clinically meaningful heterogeneity of OSA. Zinchuk and Yaggi (2020) [4] explicitly argued that “phenotypic subtypes of OSA represent a challenge and opportunity for precision medicine,” noting that AHI-based severity categories lump together patients with fundamentally different pathophysiologies, symptom profiles, and treatment responses. Importantly, this inadequacy has been demonstrated not only through clinical clustering but also through polysomnographic phenotyping: Zinchuk et al. (2018) [11] showed that PSG-derived phenotypes carry differential cardiovascular risk independently of AHI, while Kim et al. (2020) [10] confirmed in a Korean cohort that PSG-based phenotypes predict mortality independently of AHI severity. Our data provide complementary empirical support from an entirely clinical perspective, demonstrating that two populations with identical AHI severity can differ markedly in their cardiometabolic risk profile, suggesting that they may require fundamentally different therapeutic approaches, the symptomatic cluster potentially benefiting most from CPAP for symptom relief, while the asymptomatic cardiometabolic cluster may require integrated cardiovascular risk management alongside airway therapy.

This dissociation is further reinforced by the findings of Silveira et al. (2022) [7], who demonstrated in their 5-year follow-up that baseline phenotypic cluster membership predicted differential long-term outcomes (comorbidity incidence, medication use, hospitalization, and mortality) independently of AHI severity, providing direct evidence that phenotypic classification captures clinically actionable information that AHI alone cannot. Yassen et al. (2022) [18], in a landmark analysis of 20,164 patients from the European Sleep Apnea Database (ESADA), identified seven phenotypic clusters and demonstrated that the response to positive airway pressure treatment varied significantly across clusters, with two of the seven showing only limited improvement in daytime

sleepiness following PAP treatment. Their finding that treatment response is phenotype-dependent further underscores the inadequacy of AHI-based classification for therapeutic decision-making.

The notion that identical AHI values can mask fundamentally different clinical presentations is also supported by the work of Dai et al. (2023) [46], who proposed the Sleep Breathing Impairment Index (SBII) as a more nuanced alternative to AHI. Their cluster analysis revealed three groups with no significant differences in AHI between them, yet with markedly different SBII profiles and Framingham cardiovascular risk scores, providing further evidence that the phenotypic dimension captured by clinical clustering is orthogonal to the severity dimension captured by AHI.

5.2.2.2 The “morphotypic inversion” in the Spanish scientific approach

A visually and conceptually striking finding from the radar plot analysis of the Spanish scientific approach was the “morphotypic inversion” between the two clusters. As illustrated in Fig. 9, the radar polygons of Spanish Clusters 1 and 2 exhibit an immediate visual “mirror image” pattern: Cluster 1 (symptomatic, low comorbidity) presented an expanded symptomatic axis but a contracted cardiovascular axis, while Cluster 2 (asymptomatic, high comorbidity) exhibited the exact opposite an expanded cardiovascular and metabolic axis but a contracted symptomatic axis.

The two polygons are effectively geometric complements of each other, communicating at a glance that these represent two fundamentally different clinical presentations of the same disease severity. The inverse relationship between symptom severity and cardiometabolic burden may reflect distinct pathophysiological pathways: the symptomatic cluster may be driven primarily by upper airway anatomy and low arousal threshold, producing pronounced nocturnal respiratory symptoms, while the asymptomatic cluster may be driven by chronic metabolic dysfunction and vascular remodeling, producing a “silent” form of OSA with high cardiovascular risk but attenuated perception of sleep disruption.

This pattern is consistent with the “minimally symptomatic versus sleepy” dichotomy that has emerged as one of the most reproducible findings in the OSA phenotyping literature. Joosten et al. (2012) [27], in one of the earliest phenotyping studies, identified a “minimally symptomatic” phenotype in patients with mild to moderate OSA who exhibited low Epworth scores despite confirmed disease, suggesting that the symptom–comorbidity dissociation may be a fundamental feature of OSA heterogeneity rather than an artifact of our specific cohort.

5.2.3 Cross-cohort phenotypic archetypes

Integrating the results from both approaches and both cohorts, a set of recurrent phenotypic archetypes emerges that transcends the specificities of each individual clustering solution:

(i) **The “obese, hypertensive, high-severity” archetype.** This archetype is represented by Portuguese C1 (clinical approach: obese, hypertensive, insomniac, mean AHI 33.3), Spanish C3

(clinical approach: middle-aged, obese, universally hypertensive, mean AHI 56.6), and the high-comorbidity clusters of the scientific approach (Portuguese C1 and Spanish C2). Its defining features: concurrent obesity, hypertension, and elevated AHI are structurally analogous to the “multi-disease old obese” cluster (Cluster 3) of Bailly et al. (2016) [16], the “older, obese, severe OSA” cluster of Silveira et al. (2022) [7], and the “obesity-hypertension” cluster (C7) of Testelmans et al. (2022) [37]. Across these studies and our own, the co-occurrence of obesity and hypertension consistently identifies the subgroup with the most severe disease and, where longitudinal data are available, the worst cardiovascular outcomes. Tondo et al. (2024) [38] further demonstrated, using machine learning in the ESADA cohort, that phenotypes defined by high cardiometabolic burden predicted long-term mortality risk, providing prognostic validation for this archetype.

(ii) The “young, low-comorbidity” archetype. This archetype is represented by Spanish C4 (clinical approach: young, normotensive, mean AHI 34.3) and Portuguese C5 (clinical approach: overweight, normotensive, non-insomniac, mean AHI 14.9). Despite differing AHI severity levels (attributable to the referral bias discussed above), both clusters share the essential feature of minimal cardiometabolic burden. This archetype corresponds to the “young symptomatic” (Cluster 1, median age 48, few comorbidities) and “young snorer” (Cluster 4, median age 49, few comorbidities) clusters of Bailly et al. (2016) [16], and to the Cluster 2 (younger, less obese males with upper airway symptoms) identified by Venkatnarayan and Krishnaswamy (2023) [30]. The persistence of this archetype across European and Asian populations suggests a distinct pathophysiological pathway, possibly anatomically driven or related to low arousal threshold, that is independent of the cardiometabolic pathway.

(iii) The “COMISA” archetype. Uniquely prominent in our Portuguese cohort, this archetype is represented by clusters where insomnia functions as an independent stratification variable (C1, C3, C4 in the clinical approach; C1 in the scientific approach). The identification of insomnia as a phenotypic discriminator in our data adds to a small but growing body of evidence, including Wei et al. (2023) [31] and Miranda et al. (2025) [59], that COMISA-related symptoms capture a clinically meaningful dimension of OSA heterogeneity. The absence of insomnia as a discriminator in the Spanish cohort ($p > 0.999$) suggests that COMISA phenotyping may be population- or referral-dependent, and highlights the importance of including sleep quality variables in future phenotyping studies.

(iv) The “asymptomatic, high cardiometabolic burden” archetype. This archetype, most clearly represented by Spanish C2 in the scientific approach (hypertensive, dyslipidemic, low EDS, mean AHI 44.8), is clinically important because it represents patients at high cardiovascular risk who may not seek clinical attention due to the absence of subjective symptoms. This “silent” phenotype has been identified in multiple studies, including the “old obese” cluster (Cluster 2) of Bailly et al. (2016) [16], characterized by the oldest age, median BMI, few symptoms, and a high comorbidity subgroup. As previously noted, the treatment-resistant clusters identified by Yassen et al. (2022) [18] align conceptually with this asymptomatic archetype, suggesting that these patients’ disease manifestation is primarily cardiovascular rather than symptom-driven and who may therefore require therapeutic approaches beyond CPAP alone. The identification of this

archetype in our data underscores the clinical imperative for proactive cardiovascular screening in OSA patients who present without excessive daytime sleepiness.

5.3 Confrontation of clustering approaches

The dual-pathway methodology applied in this study provides critical insights into how variable selection shapes phenotypic discovery. The clinical approach, methodologically analogous to the filter-based feature selection employed by Ferreira-Santos and Rodrigues (2021) [6], forces the algorithm to partition the cohort strictly along axes that correlate with disease presence. This supervised pre-filtering predictably yields AHI-stratified clusters with high clinical interpretability. Conversely, the data-driven scientific approach, by avoiding diagnostic pre-filtering, allows broader behavioral and pathophysiological axes to emerge, most notably exposing the orthogonality between clinical presentation and AHI severity discussed in Section 5.2.2.

Rather than representing conflicting results, these complementary views serve as a form of internal validation. As noted by the systematic review of Almeida et al. (2025) [22], most phenotyping studies employ a single variable selection strategy, limiting the ability to assess robustness. By applying both strategies to the same datasets, our study demonstrates that the core phenotypic axes emerge consistently regardless of the analytical pathway. Furthermore, the identification of identical cluster numbers ($k = 2$) under the scientific approach in both cohorts provides additional evidence of the stability of the coarser phenotypic partition. This methodological triangulation strengthens the confidence in the identified phenotypic structures and addresses a critical limitation identified in the OSA phenotyping literature.

From a methodological standpoint, Almeida et al. (2025) [22] identified hierarchical clustering and k-means as the dominant algorithms in OSA phenotyping (59% and 27% of 38 studies, respectively). Our use of k-modes [23], which extends k-means to categorical data, addresses a limitation highlighted in the review: the inappropriate application of distance-based algorithms to predominantly categorical datasets. Several studies in the Almeida et al. review applied k-means or hierarchical clustering with Euclidean distance to mixed or categorical variables, which can introduce artifactual cluster structures. Our exclusive use of k-modes with the simple matching dissimilarity measure ensures that the clustering is mathematically appropriate for the data type, a methodological advantage shared only with the earlier work of Ferreira-Santos and Rodrigues (2018, 2023) [25; 26]. Ghorvei et al. (2024) [48], in a systematic comparison of unsupervised machine-learning methods applied to PSG-related phenotyping, demonstrated that different clustering algorithms yield substantially different phenotypic solutions from the same dataset, underscoring that the choice of algorithm is not neutral and further justifying our dual-approach validation strategy. Our study design also mitigates common sources of bias identified by the QUADAS-2 analysis in the Almeida et al. review, specifically by: (i) applying clustering to the full cohort rather than restricting to diagnosed patients; and (ii) using AHI as a post-hoc validation metric rather than an input variable.

Beyond algorithmic considerations, the two visualization strategies employed in this study served complementary analytical functions. While the heatmap representations facilitated detailed variable-by-variable comparison across clusters, the radar plots proved particularly valuable for cross-cohort phenotypic comparison, as demonstrated by the morphotypic inversion discussed in Section 5.2.2. From a clinical translation perspective, the geometric signature captured by radar plots is well suited for bedside decision-making: a clinician can rapidly assess whether a patient’s profile aligns with a given phenotype by comparing the patient’s individual radar footprint to the cluster template, in a manner analogous to electrocardiographic pattern recognition in cardiology [24].

5.4 Limitations

Several limitations must be acknowledged when interpreting the findings of this study. Both cohorts were drawn from single tertiary centers, which limits generalizability to primary care or community-based populations. Ntenta et al. (2022) [36] demonstrated that OSA phenotypes can differ substantially between tertiary and primary care settings, suggesting that the phenotypic structures identified in our hospital-based populations may not fully represent the phenotypic diversity of OSA in the broader community. The referral bias discussed in Section 5.1 further constrains the external validity of the Portuguese results, as the in-lab PSG subset represents a highly selected subpopulation.

An additional concern relates to the substantial sample size imbalance between the two cohorts: the Spanish dataset ($n = 355$) is nearly twice the size of the Portuguese dataset ($n = 188$). This asymmetry has implications for clustering stability and statistical power. Larger samples provide a richer representation of population heterogeneity, enabling the identification of rarer phenotypic subgroups, whereas smaller samples may yield less stable cluster assignments and are more susceptible to the influence of outliers. The coarser cluster solutions obtained under the data-driven approach in the Portuguese cohort (two clusters, as in Spain) may partly reflect this reduced sample size, which limits the algorithm’s capacity to resolve finer-grained phenotypic structure. By comparison, the largest phenotyping studies: Bailly et al. (2016) [16] with 18,263 patients and Yassen et al. (2022) [18] with 20,164 patients, identified six and seven clusters, respectively, suggesting that sample size is a key determinant of achievable phenotypic resolution.

The retrospective nature of data collection also introduces potential inconsistencies in variable documentation across clinical teams and time periods. A key strength of the present study is its reliance on readily accessible clinical, demographic, and comorbidity data, maximizing the translational value of the identified phenotypes to routine clinical practice. To appropriately handle the categorical nature of these standard clinical variables, the analysis was deliberately restricted to the k-modes algorithm. Consequently, specialized continuous metrics such as high-resolution polysomnographic indices (e.g., hypoxic burden), sleep microarchitecture variables, or continuous cardiovascular biomarkers were not included in the clustering framework. While this pragmatic design choice successfully delineated actionable clinical phenotypes without requiring complex

or invasive diagnostic data, it inherently precludes the identification of anatomically or physiologically driven phenotypes [6; 16; 32]. Future research could aim to bridge these approaches by integrating standard clinical categorical profiles with advanced anatomical markers.

From an algorithmic perspective, the k-modes algorithm, while appropriate for categorical data, produces hard cluster assignments and does not account for the possibility that patients may belong to multiple phenotypes simultaneously. Fuzzy or probabilistic clustering methods could offer a more nuanced representation of phenotypic overlap [48]. It should also be noted that the absence of longitudinal follow-up data prevents assessment of the predictive validity of the identified phenotypes, namely whether cluster membership predicts differential treatment response, cardiovascular outcomes, or mortality. Silveira et al. (2022) [7] demonstrated that such predictive validation is feasible and clinically meaningful, having shown that baseline cluster membership predicted 5-year mortality risk in their Spanish cohort.

Chapter 6

Conclusion

This dissertation addressed the fundamental challenge of phenotypic heterogeneity in OSA by exploring the utility of categorical data-driven clustering, evaluating distinct variable selection strategies, and employing multidimensional visualization techniques applied exclusively to pre-triage clinical variables across two independent Iberian cohorts. The pervasive reliance on the AHI as the sole arbiter of disease severity has historically constrained clinical decision-making, often leading to a monolithic approach that fails to capture the true multidimensional nature of the disorder. By integrating clinical, demographic, and behavioral data from two independent and geographically distinct cohorts, this study successfully demonstrated that clinically meaningful OSA phenotypes can be identified before patients undergo resource-intensive polysomnography.

The methodological framework centered on the k-modes algorithm, a clustering technique specifically designed to handle categorical data, thereby avoiding the algorithmic distortions introduced when traditional continuous distance metrics are applied to clinical variables. A defining methodological feature of this work was the implementation of a dual variable selection strategy. The clinically-guided approach, which pre-filtered variables based on their univariate association with OSA diagnosis, yielded highly granular and interpretable partitions. In the Portuguese cohort, it revealed a severe phenotype characterized by obesity, hypertension, and insomnia, whereas in the Spanish cohort, it identified a middle-aged, obese, and hypertensive group as the most severely affected. In contrast, the scientific approach, which incorporated all available variables without diagnostic supervision, exposed a broader axis of clinical variance. Notably, it demonstrated that two patient clusters could exhibit markedly different clinical profiles (e.g., a highly symptomatic but normotensive group versus an asymptomatic but highly comorbid group) while presenting virtually identical AHI distributions, a dissociation that empirically underscores the critical limitation of AHI-based classification and highlights the added value of multidimensional phenotyping.

Furthermore, this dissertation employed radar plots alongside traditional heatmaps for the visualization of clinical clusters. This dual visualization strategy proved to be highly complementary: while heatmaps are effective for analyzing population-level variable distributions, radar plots offer an intuitive geometric signature of each phenotype, facilitating rapid, gestalt recognition of clinical profiles that is highly translatable to point-of-care decision-making.

Notwithstanding the limitations detailed in Chapter 5, namely the retrospective design, the reliance on single-center registries for each cohort, cohort size imbalance, and the absence of longitudinal outcome data, the findings provide a robust foundation for future work centered on external validation, algorithmic triangulation, and the incorporation of longitudinal outcome data to transition these descriptive phenotypes into clinically actionable prediction tools.

In conclusion, this research confirms that pre-triage clinical phenotyping is both feasible and highly informative. By moving beyond the single severity dimension of the AHI, the application

of appropriate categorical clustering techniques to routine clinical data provides a robust foundation for precision medicine in sleep apnea. This multidimensional approach promises to optimize diagnostic triage, personalize therapeutic interventions, and ultimately reduce the clinical and socioeconomic burden associated with this prevalent disorder.

6.1 Future directions

Several avenues for future research emerge from the present findings. The exclusive reliance on k-modes, while justified for categorical datasets, should be complemented by alternative clustering algorithms to assess the stability of the identified phenotypes across methods. Approaches such as Latent Class Analysis (LCA) [48] or modified hierarchical clustering with appropriate categorical distance metrics would provide methodological triangulation and strengthen the evidence for phenotypic robustness without violating the assumptions of the data types. The combined unsupervised-supervised framework proposed by Ma et al. (2021) [40], which uses supervised learning to identify key features of high-risk phenotypes after unsupervised clustering, represents a particularly promising direction for translating phenotypic structures into clinically actionable prediction tools.

Equally important is the need for prospective studies incorporating longitudinal outcome data (treatment response, cardiovascular events, mortality) to transition from descriptive phenotyping to clinically actionable patient stratification. The work of Silveira et al. (2022) [7], which demonstrated that baseline phenotypes predict 5-year outcomes, provides a methodological template for such studies. Beyond treatment efficacy, phenotype-guided prediction of treatment adherence represents a complementary avenue: Hamada et al. (2024) [39] used hierarchical clustering to predict CPAP adherence patterns from baseline clinical profiles, demonstrating that phenotypic classification can inform not only which treatment to prescribe but also whether a patient is likely to adhere to it.

The finding that insomnia functions as an independent stratification variable in the Portuguese cohort opens an important avenue for COMISA-focused phenotyping. Future studies should systematically incorporate categorical validated insomnia measures (e.g., the Insomnia Severity Index clinical categories) alongside traditional OSA variables to determine whether COMISA phenotypes exhibit differential treatment responses, as suggested by the growing evidence that combined CBT-I and CPAP approaches may outperform either intervention alone in COMISA patients [59].

Ultimately, external validation of the identified phenotypes in independent, larger, and more geographically diverse cohorts remains a critical step to establish generalizability and support clinical implementation. The cross-cohort convergence observed in our scientific approach provides an encouraging preliminary signal of phenotypic robustness, but multicenter studies spanning multiple European countries and healthcare systems are needed to confirm that the identified archetypes are truly generalizable rather than artifacts of the specific Iberian context.

References

- [1] Adam V Benjafield, Najib T Ayas, Peter R Eastwood, et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. *The Lancet Respiratory Medicine*, 7(8):687–698, 2019.
- [2] G. Iannella, A. Pace, M. G. Bellizzi, et al. The global burden of obstructive sleep apnea. *Diagnostics*, 15(9):1088, 2025.
- [3] E. Fontanilles Arbones, N. Salord Oleo, et al. Phenotypes of obstructive sleep apnea in women: a real-life cohort study. *Sleep Medicine*, 121:295–302, 2024.
- [4] Andrey Zinchuk and H. Klar Yaggi. Phenotypic subtypes of OSA: a challenge and opportunity for precision medicine. *Chest*, 157(2):403–420, 2020.
- [5] V. Luzzi, M. Mazur, M. Guaragna, et al. Correlations of obstructive sleep apnea syndrome and daytime sleepiness with the risk of car accidents in adult working population: a systematic review and meta-analysis. *Journal of Clinical Medicine*, 11(14):3971, 2022.
- [6] Daniela Ferreira-Santos and Pedro Pereira Rodrigues. Enhancing obstructive sleep apnea diagnosis with screening through disease phenotypes: algorithm development and validation. *JMIR Medical Informatics*, 9(6):e28230, 2021.
- [7] María Guadalupe Silveira, Gabriel Sampol, Miriam Mota-Foix, et al. Cluster-derived obstructive sleep apnea phenotypes and outcomes at 5-year follow-up. *Journal of Clinical Sleep Medicine*, 18(2):597–607, 2022.
- [8] United Nations. Transforming our world: the 2030 agenda for sustainable development, 2015. General Assembly Resolution 70/1.
- [9] Hui-Ling Tan, Kelvin Kian Loong Yang, and Amanda Kah Jia Lim. Anatomy and pathophysiology of upper airway obstructive sleep apnoea: review of the current literature. *Sleep Medicine Research*, 12(2):61–67, 2021. Open Access, CC BY-NC 4.0. DOI: 10.17241/smr.2021.00941.
- [10] Jeong-Whun Kim, Tae-Bin Won, Chae-Seo Rhee, et al. Polysomnographic phenotyping of obstructive sleep apnea and its implications in mortality in Korea. *Scientific Reports*, 10(1):13207, 2020.
- [11] Andrey V. Zinchuk, S. Jeon, B. B. Koo, et al. Polysomnographic phenotypes and their cardiovascular implications in obstructive sleep apnoea. *Thorax*, 73(5):472–480, 2018.
- [12] Yutaka Kozu, Yusuke Kurosawa, Shiho Yamada, et al. Cluster analysis identifies a pathophysiologically distinct subpopulation with increased serum leptin levels and severe obstructive sleep apnea. *Sleep and Breathing*, 25(2):767–776, 2021.
- [13] Soonhyun Yook, H. R. Park, D. Seo, et al. Obstructive sleep apnea subtyping based on apnea and hypopnea specific hypoxic burden is associated with brain aging and cardiometabolic syndrome. *Computers in Biology and Medicine*, 185, 2025.
- [14] American Academy of Sleep Medicine. *International Classification of Sleep Disorders*. American Academy of Sleep Medicine, Darien, IL, 3rd edition, 2014.

- [15] Richard B Berry, Rita Brooks, Charlene E Gamaldo, et al. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, Version 2.4*. American Academy of Sleep Medicine, Darien, IL, 2017.
- [16] S. Bailly, M. Destors, Y. Grillet, et al. Obstructive sleep apnea: a cluster analysis at time of diagnosis. *PloS One*, 11(6):e0157318, 2016.
- [17] Brendan T Keenan, Jihye Kim, Bharati Singh, et al. Recognizable clinical subtypes of obstructive sleep apnea across international sleep centers: a cluster analysis. *Sleep*, 41(3):zsx214, 2018.
- [18] A. Yassen, K. Coboeken, S. Bailly, et al. Baseline clusters and the response to positive airway pressure treatment in obstructive sleep apnoea patients: longitudinal data from the European Sleep Apnea Database cohort. *ERJ Open Research*, 8(4):00132–2022, 2022.
- [19] Hideaki Nakayama, Mina Kobayashi, Satoru Tsuiki, et al. Obstructive sleep apnea phenotypes in men based on characteristics of respiratory events during polysomnography. *Sleep and Breathing*, 23(4):1087–1094, 2019.
- [20] Mireia Gasa, N. Salord, E. Fontanilles, et al. Polysomnographic phenotypes of obstructive sleep apnea in a real-life cohort: a pathophysiological approach. *Archivos de Bronconeumología*, 59(10):638–644, 2023.
- [21] Dirk Pevernagie. Future treatment of sleep disorders: syndromic approach versus management of treatable traits? *Sleep Medicine Clinics*, 16:465–473, 2021.
- [22] Francisca Almeida, Adrián Martín-Montero, Gonzalo C. Gutiérrez-Tobal, Javier Gómez-Pilar, Roberto Hornero, Pedro Pereira Rodrigues, and Daniela Ferreira-Santos. The potential of clustering methods for pre-test triage in sleep medicine: a systematic review. *Manuscript*, 2025. Systematic review registered on PROSPERO (CRD42025620166).
- [23] Zhexue Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2:283–304, 1998.
- [24] Daniela Ferreira-Santos and Pedro Pereira Rodrigues. Phenotyping obstructive sleep apnea patients: a first approach to cluster visualization. In *Studies in Health Technology and Informatics*, pages 75–79. IOS Press, 2018.
- [25] Daniela Ferreira-Santos and Pedro Pereira Rodrigues. Finding groups in obstructive sleep apnea patients: a categorical cluster analysis. In *Proceedings – IEEE Symposium on Computer-Based Medical Systems*, pages 387–392. IEEE, 2018.
- [26] Daniela Ferreira-Santos and Pedro Pereira Rodrigues. Obstructive sleep apnea: a categorical cluster analysis and visualization. *Pulmonology*, 29(3):207–213, 2023.
- [27] Simon A. Joosten, Kaïs Hamza, Scott Sands, et al. Phenotypes of patients with mild to moderate obstructive sleep apnoea as confirmed by cluster analysis. *Respirology*, 17(1):99–107, 2012.
- [28] S. Romero-Peralta, F. García-Río, et al. Defining the heterogeneity of sleep apnea syndrome: a cluster analysis with implications for patient management. *Archivos de Bronconeumología*, 58(2):125–134, 2022.

- [29] Xiao Lei Zhang, Li Zhang, Yi Ming Li, et al. Multidimensional assessment and cluster analysis for OSA phenotyping. *Journal of Clinical Sleep Medicine*, 18(7):1779–1788, 2022.
- [30] Kavitha Venkatnarayan, Uma Maheswari Krishnaswamy, et al. Identifying phenotypes of obstructive sleep apnea using cluster analysis. *Sleep and Breathing*, 27(3):879–886, 2023.
- [31] Huasheng Wei, Jie Zhu, Fei Lei, et al. Clinical phenotypes of obstructive sleep apnea: a cluster analysis based on sleep perception and sleep quality. *Sleep and Breathing*, 27(5):1829–1837, 2023.
- [32] Liping Huang, Ying Xu, Xu Gong, and Xuemei Gao. Anatomical phenotype of obstructive sleep apnea patients based on cluster analysis. *Orthodontics & Craniofacial Research*, 26(4):608–617, 2023.
- [33] H. J. An, S. H. Baek, S. W. Kim, et al. Clustering-based characterization of clinical phenotypes in obstructive sleep apnoea using severity, obesity, and craniofacial pattern. *European Journal of Orthodontics*, 42(1):93–100, 2020.
- [34] Su-Jung Kim, Waleed Maqbul Alnakhli, et al. Multi-perspective clustering of obstructive sleep apnea towards precision therapeutic decision including craniofacial intervention. *Sleep and Breathing*, 25(1):85–94, 2020.
- [35] George D. Vavougiou, G. Natsios, et al. Phenotypes of comorbidity in OSAS patients: combining categorical principal component analysis with cluster analysis. *Journal of Sleep Research*, 25(1):31–38, 2016.
- [36] P. K. Ntenta, G. D. Vavougiou, et al. Obstructive sleep apnea syndrome comorbidity phenotypes in primary health care patients in Northern Greece. *Healthcare*, 10(2):338, 2022.
- [37] Dries Testelmans, M. A. Spruit, B. Vrijsen, et al. Comorbidity clusters in patients with moderate-to-severe OSA. *Sleep and Breathing*, 26(1):195–204, 2022.
- [38] Pasquale Tondo, G. Scioscia, S. Bailly, et al. Exploring phenotypes to improve long-term mortality risk stratification in obstructive sleep apnea through a machine learning approach. *European Journal of Internal Medicine*, 2024.
- [39] E. Hamada, M. Yamauchi, Y. Fujita, et al. Multidimensional prediction of continuous positive airway pressure adherence. *Sleep Medicine*, 122:177–184, 2024.
- [40] Eun-Yeol Ma, Jeong-Whun Kim, Youngmin Lee, et al. Combined unsupervised-supervised machine learning for phenotyping complex diseases with its application to obstructive sleep apnea. *Scientific Reports*, 11(1):4457, 2021.
- [41] Son T. Mai, Sihem Amer-Yahia, S. Bailly, et al. Evolutionary active constrained clustering for obstructive sleep apnea analysis. *Data Science and Engineering*, 3(4):359–378, 2018.
- [42] Donato Lacedonia, Giovanna Elisiana Carpagano, et al. Characterization of obstructive sleep apnea–hypopnea syndrome (OSA) population by means of cluster analysis. *Journal of Sleep Research*, 25(6):724–730, 2016.
- [43] Takashi Oki, Yutaka Kozu, et al. The relationship between leptin levels and continuous positive airway pressure treatment: a cluster analysis. *Sleep Science*, 17(2):e143–e150, 2024.

- [44] Hitomi Ida, Tatsuo Suga, Masaharu Nishimura, and Yasuhiro Aoki. Unique clinical phenotypes of patients with obstructive sleep apnea in a Japanese population: a cluster analysis. *Journal of Clinical Sleep Medicine*, 18(3):895–902, 2022.
- [45] H. J. Kim, J. W. Choi, E. J. Joo, et al. Identifying obstructive sleep apnea subtypes and its correlation with mood states. *Chronobiology in Medicine*, 3(3):120–125, 2021.
- [46] Lu Dai, Wenhao Cao, Jinmei Luo, et al. The effectiveness of sleep breathing impairment index in assessing obstructive sleep apnea severity. *Journal of Clinical Sleep Medicine*, 19(2):267–274, 2023.
- [47] M. Tsuchiya, A. A. Lowe, E. K. Pae, and J. A. Fleetham. Obstructive sleep apnea subtypes by cluster analysis. *American Journal of Orthodontics and Dentofacial Orthopedics*, 101(6):533–542, 1992.
- [48] M. Ghorvei, T. Karhu, S. Hietakoste, D. Ferreira-Santos, et al. A comparative analysis of unsupervised machine-learning methods in PSG-related phenotyping. *Journal of Sleep Research*, 34(3):e14349, 2024.
- [49] Stefan Mihaicuta, Mihai Udrescu, Alexandru Topirceanu, and Lucretia Udrescu. Network science meets respiratory medicine for OSAS phenotyping and severity prediction. *PeerJ*, 5:e3289, 2017.
- [50] Javier Gómez-Pilar, Daniela Ferreira-Santos, Pedro Pereira-Rodrigues, David Gozal, Roberto Hornero, and Gonzalo C. Gutiérrez-Tobal. Subject-based association networks for clustering analysis and phenotyping of pediatric sleep apnea. In *Proceedings CASEIB*, 2022.
- [51] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.
- [52] Alexandru Topîrceanu, Lucretia Udrescu, Mihai Udrescu, and Stefan Mihaicuta. Gender phenotyping of patients with obstructive sleep apnea syndrome using a network science approach. *Journal of Clinical Medicine*, 9(12):4025, 2020.
- [53] Gonzalo C. Gutiérrez-Tobal et al. Pediatric sleep apnea: the overnight electroencephalogram as a phenotypic biomarker. *Frontiers in Neuroscience*, 15:644697, 2021.
- [54] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS One*, 9(6):e98679, 2014.
- [55] Javier Gómez-Pilar et al. SpO2-based weighted correlation networks for adult OSA phenotyping. In *47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2025.
- [56] Richard B Berry, Rita Brooks, Charlene Gamaldo, Susan M Harding, Carole L Marcus, Bradley V Vaughn, et al. *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications*. American Academy of Sleep Medicine, Darien, IL, 2012.
- [57] A. Agresti. *Categorical Data Analysis*. John Wiley & Sons, New York, 2nd edition, 2002.

- [58] Nelis J. de Vos. kmodes categorical clustering library. <https://github.com/nicodv/kmodes>, 2015–2024.
- [59] M. Miranda, D. Ferreira-Santos, D. Alvarez-Estevez, and P. P. Rodrigues. Characterization of COMISA, insomnia, and OSA profiles using polysomnographic and heart rate variability data: a two-population study. *IEEE Access*, 2025. In press.
- [60] D. F. Ransohoff and A. R. Feinstein. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *New England Journal of Medicine*, 299(17):926–930, 1978.