# A Foundation Model for Sleep-Based Risk Stratification and Clinical Outcomes

**Jeffrey Rogers**

jeffrogers@us.ibm.com

IBM Research

**Erhan Bilal**

Digital Health, IBM Research https://orcid.org/0009-0002-0778-6635

**Matheus Lima Araujo**

Sleep Disorders Center, Cleveland Clinic Foundation

**Kristen Beck**

Digital Health, IBM Research

**Catherine Heinzinger**

Sleep Disorders Center, Cleveland Clinic Foundation

**Samer Ghosn**

Department of Biomedical Engineering, Cleveland Clinic Foundation

**Carl Saab**

Cleveland Clinic https://orcid.org/0000-0003-4665-9946

**Nancy Schaefer**

Sleep Disorders & Epilepsy Centers, Cleveland Clinic Foundation

**Reena Mehra**

University of Washington https://orcid.org/0000-0002-6222-2675

**Article**

**Keywords:**

**Posted Date:** April 10th, 2025

**DOI:** https://doi.org/10.21203/rs.3.rs-6307069/v1



**Additional Declarations:**
There is **NO** Competing Interest.

Note: Supplementary Tables 2, 3, 6, 10, and 11 are not available with this version.

## TITLE

A Foundation Model for Sleep-Based Risk Stratification and Clinical Outcomes


## ABSTRACT

Clinical diagnosis of sleep disorders, which are recognized contributors to morbidity and mortality, often relies on polysomnography (PSG) data. However, the vast physiologic data collected during PSG is underutilized, presenting a key opportunity to enhance characterization of sleep dysfunction and predict clinical outcomes. We introduce a sleep foundation model that uniquely integrates PSG time-series signals and electronic medical record data. Using a diverse dataset (n=10,000; mean observation period 14.5±7.1 years), our transformer-based model generates data-driven representations of latent physiological patterns. When clustered, we identified subpopulations with differential health trajectories. The highest risk-group exhibited strong correlations with all-cause mortality (unadjusted hazard ratio [HR] 4.83, 95% confidence interval [CI] 3.60–6.50, p<0.001) as well as cardiovascular outcomes and neurological outcomes, even after accounting for traditional measures. External validation in a National Sleep Research Resource cohort confirmed findings. We created a novel, clinically applicable framework leveraging information-dense PSG data to inform risk stratification and predict health outcomes beyond traditional methods.


## AUTHORS


Erhan Bilal
Digital Health, IBM Research, T.J. Watson Research Center, Yorktown Heights, USA

Matheus Lima Diniz Araujo
Sleep Disorders Center, Neurological Institute Cleveland Clinic Foundation, Cleveland, USA

Kristen L. Beck
Digital Health, IBM Research, IBM Research Almaden Lab, San Jose, USA

Catherine M. Heinzinger
Sleep Disorders Center, Neurological Institute Cleveland Clinic Foundation, Cleveland, USA

Samer Ghosn
Department of Biomedical Engineering, Cleveland Clinic Foundation, Cleveland, USA

Carl Y. Saab
Department of Biomedical Engineering, Cleveland Clinic Foundation, Cleveland, USA
School of Medicine, Case Western Reserve University, Cleveland, USA
Department of Engineering, Brown University, Providence, USA

Nancy Foldvary Schaefer
Sleep Disorders & Epilepsy Centers, Neurological Institute Cleveland Clinic Foundation, Cleveland, USA

Jeffrey L. Rogers
Digital Health, IBM Research, T.J. Watson Research Center, Yorktown Heights, USA

Reena Mehra
Division of Pulmonary, Critical Care, and Sleep Medicine, University of Washington, Seattle, USA

## COMPETING INTERESTS

No competing interest declared.

## DATA AVAILABILITY

Data are available upon request and can be shared in accordance with Institutional Review Board and Data User Agreement limitations.

## CODE AVAILABILITY

Code is available upon request and can be shared in accordance with intellectual property and copyright limitations.

## INTRODUCTION

Sleep is fundamental to health and well-being. Conversely, sleep disorders— associated with a range of physical, psychiatric, and cognitive symptoms— increase the risk of cardiovascular, metabolic, respiratory, and neurological conditions, as well as mortality. Reliable and accurate assessment of sleep and its associated clinical outcomes in a comprehensive, data-driven manner is a challenge, mainly due to the number of different disorders and their variable effects on general health, co-morbidities, and genetic predisposition of an individual.

Diagnosing sleep disorders requires a comprehensive sleep and medical history supplemented by validated, self-assessment instruments and sleep diary, physical examination, and, in many cases— such as sleep disordered breathing (SDB) or central disorders of hypersomnolence— a diagnostic sleep study. A laboratory-based overnight sleep study, known as a polysomnogram (PSG), is conducted by a trained technologist and involves direct, real-time, continuous physiologic monitoring of electroencephalography (EEG), electrooculography (EOG), chin and limb electromyography (EMG), electrocardiography (EKG), oxygen saturation, oral and nasal airflow, respiratory effort (inductance plethysmography), position, snoring, video, and, in some cases, carbon dioxide monitoring. These rich and complex multimodal physiological parameters are collected over the sleep period which typically range from six to eight hours and are later interpreted by a physician. Yet current diagnostic criterion for SDB, as an example, is based solely on the apnea-hypopnea index (AHI)[1], an essentially unidimensional measure of the frequency of upper airway closure. Hence, the

richness and complexity of sleep physiological data are not efficiently and systematically incorporated into the ascertainment of clinical diagnoses or decision-making.

Several clinic-based and epidemiologic studies have examined the utility of PSG-based clustering of aggregate indices such as the AHI, arousal index, and sleep stages scored and annotated by technologists to serve as predictors of clinical outcomes[2-4]. The fidelity of this approach, however, is limited due to its reliance on coarse aggregate PSG measures, which depend on individual technologist interpretations rather than a deeper, more direct approach that leverages the information contained in the continuous multimodal physiologic dataset. Furthermore, in other areas of medicine, machine learning (ML), particularly clustering techniques, has revealed hidden disease subtypes, improving risk stratification and personalized treatment. For example, unsupervised learning has identified distinct metabolic subtypes of diabetes, refining treatment beyond the standard Type 1 and Type 2 classifications[5]. Similarly, clustering techniques applied to neuroimaging and genomic data have uncovered Alzheimer's disease subgroups with varying disease trajectories and treatment responses[6]. However, traditional ML models often rely on predefined feature selection and task-specific architectures, limiting their ability to generalize across populations and adapt to new data.

To address these limitations, we introduce a foundation model for sleep health, designed to uncover latent physiological patterns associated with long-term disease development. Unlike conventional approaches, foundation models process large-scale, high-dimensional data without requiring extensive manual feature engineering, making them more scalable and generalizable. In this study, we leveraged 10,000 high-resolution PSG studies conducted at the Cleveland Clinic Sleep Disorders Center, paired with detailed longitudinal electronic medical records (EMR) spanning up to a decade, to train a transformer-based model that generates high-dimensional representations (i.e. embeddings) of a full night of sleep[7]. The model was trained specifically to ensure that these representations capture the intricate relationships between sleep architecture, respiratory events, and key physiological parameters.

Using clustering techniques, we identified distinct risk groups with significant differential longitudinal health trajectories. To validate our approach used in this clinical cohort, we examined the highest identified risk group in the multicenter, geographically diverse NIH/NHLBI Sleep Heart Health Study (SHHS)[8] population-based cohort from the National Sleep Research Resource,[9] confirming its strong association with adverse clinical outcomes. Our results demonstrate that learned sleep embeddings can effectively stratify patients into distinct risk groups, each characterized by unique disease-specific signatures that predict poor health outcomes, including cardiovascular disease, neurological disorders, and all-cause mortality. This work established a scalable, automated framework for labeling and analyzing PSG, introducing an independent risk classification schema that advances personalized sleep medicine beyond conventional methods.

# RESULTS

## Approach Overview

In this work, we sought to develop an automated time-series approach leveraging raw PSG data to generate more comprehensive patient characterizations than traditional, summary-based measures (e.g., AHI). Specifically, our study follows a three-step process: representation learning, patient stratification, and outcome analysis. First, we train a transformer-based foundation model on raw PSG time-series data, incorporating technologist annotations for sleep staging and respiratory events. This model learns high-dimensional representations (i.e. embeddings) that capture the intricate physiological patterns underlying sleep.

Next, we use these embeddings to identify distinct patient subgroups through unsupervised clustering, revealing patterns that are not apparent from conventional PSG metrics. Finally, we evaluate the clinical relevance of these subgroups by analyzing their longitudinal health trajectories, including associations with cardiovascular and neurological conditions, as well as all-cause mortality. The following sections provide detailed methodological advancements, validation, and clinical implications of this approach.

## Sleep Foundation Model

### Foundation Model Architecture

A transformer-based large language model was adapted for time-series data by replacing its original word embedding layer with a linear projection layer for time-series patches, following a strategy similar to Zhou et al[10]. More specifically, a pre-trained RoBERTa-based model[11] was fine-tuned on PSG data, employing the architecture illustrated in *Figure 1*. Unlike the original approach, where some weights remained frozen, all previously trained RoBERTa weights were fine-tuned alongside the newly introduced projection layers and parameters. This adaptation refined the model's generalized capabilities, tailoring them to the distinctive characteristics of PSG data, including waveforms, time-series structures, and multimodal sensor channels.

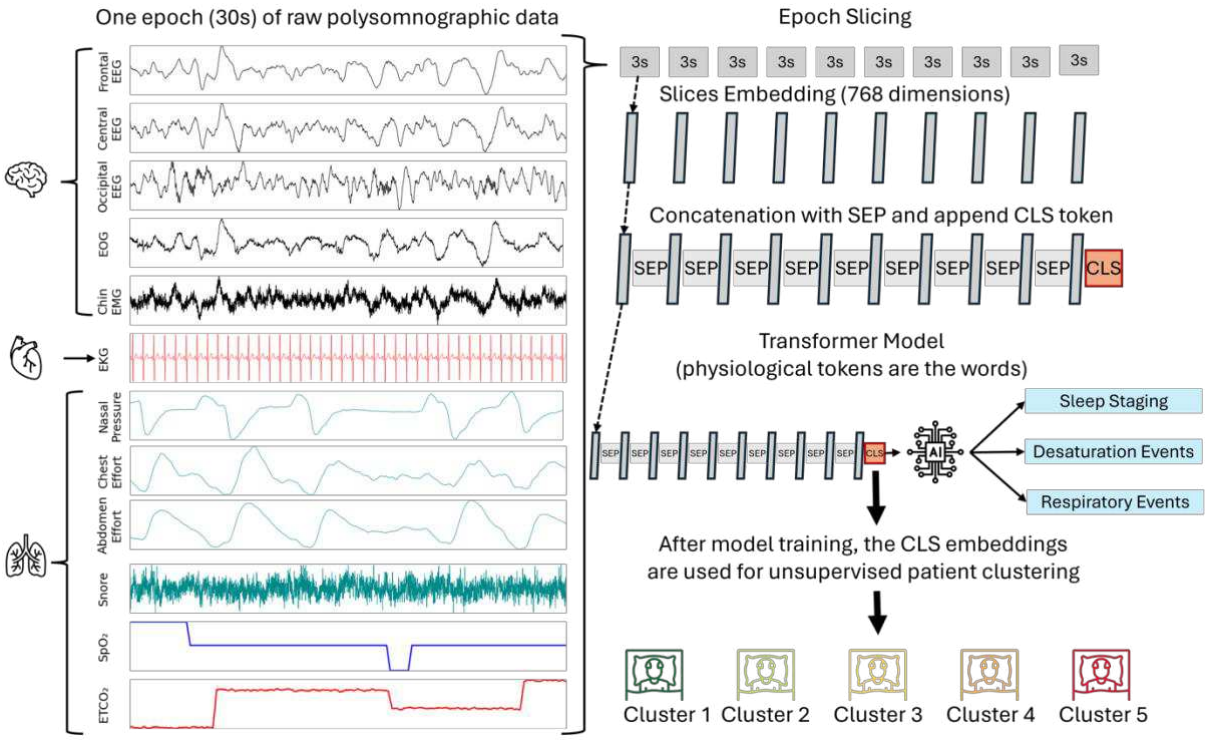



**Figure 1. Overall architecture overview.** Each 30-second PSG segment is sliced into ten nonoverlapping 3-second segments, which are projected onto a 768-dimensional space using a linear layer. Tokens from all channels are concatenated, split by separator (SEP) tokens, and a classification (CLS) token is appended at the end. The resulting sequence is passed through a Transformer backbone. The CLS embedding output is used to predict technologist-labeled sleep stages, respiratory events, and desaturations, optimized with a cross-entropy loss function. EEG: electroencephalography; EKG: electrocardiography; EMG: electromyography, $SpO_2$: oxygen saturation; $EtCO_2$: end tidal carbon dioxide.

To train the model effectively, we used PSG data that captures a wide range of physiological signals during sleep. The dataset included the following channels: C4 (EEG signal from the right central region), F4 (EEG signal from the right frontal region), O2-M2 (EEG signal from the right occipital region), $SpO_2$ (oxygen saturation measured via pulse oximetry), EKG (electrocardiogram for electrical heart signal), E1 and E2 (electrooculogram signals from left and right eyes), chin EMG (electromyogram signal from the submental muscle), nasal pressure (airflow measurement via nasal pressure), AIRFLOW (thermal sensor measuring oral airflow), chest and abdomen effort (respiratory effort signals from chest and abdominal movements via inductance plethysmography), snore (vibration sensor measuring snoring activity), and $EtCO_2$ (end-tidal carbon dioxide concentration). Each 30-second PSG epoch was divided into ten non-overlapping 3-second segments, which were linearly projected into a 768-dimensional space, resulting in 10 tokens per channel (*Figure 1*). This time segmentation was selected to mirror the time frame used by clinical technologists. These channels collectively capture neural activity, ocular movements, respiratory patterns, cardiac signals,

muscle activity, snoring, oxygen saturation, and carbon dioxide monitoring providing a comprehensive physiological snapshot during sleep.

One advantage of using the CLS token in this architecture (*Figure 1*) is that it enables the model to condense the complex multi-channel dynamics and relationships of an entire night of sleep into a compact representation while retaining the richness of the physiological data. Specifically, tokens from each channel were concatenated into a single sequence, with a learnable separator token (SEP) and a learnable CLS token appended at the end of the sequence. The CLS token served as a global representation of the input sequence for downstream classification tasks. The sequence was passed through the RoBERTa-based transformer backbone, with the CLS token embedding projected onto separate linear layers for each task: sleep stage classification, respiratory event detection, and desaturation detection.

### Model Classification Performance

Following training, model performance was evaluated on the testing subset for three classification tasks: 1. sleep stage recognition (non-rapid eye movement sleep stage 1 (N1), N2, N3, and rapid eye movement sleep stage (REM)), 2. respiratory event detection (apnea or hypopnea defined by 3% desaturation or arousal or only 4% desaturation depending on insurance requirements, as this was a clinical sample), and 3. desaturation event detection.

To account for class imbalances, the reported metrics include F1 score and average precision (AP). Both macro- and micro-averaged metrics are provided for the multi-class sleep stage classification. This classification task achieved a macro-averaged F1 of 0.75, a micro-averaged F1 of 0.86, and macro- and micro-averaged AP of 0.62 and 0.76, respectively. These metrics indicate robust performance in the automated, AI-driven labeling of sleep stages and are consistent with previously reported findings from studies that performed sleep stage classification using PSG data, achieving similar levels of performance[12].

In comparison, respiratory and desaturation classification performance results were slightly lower. Respiratory event classification yielded an F1 score of 0.65 and AP of 0.72, while desaturation event classification exhibited an F1 score of 0.59 and AP of 0.69. Minor differences in F1 and AP scores for these binary tasks likely arise from the labeling approach, wherein each 30-second epoch was labeled as a respiratory or oxygen desaturation event if at least five seconds of the manually labeled annotation overlapped with that epoch. This criterion sometimes caused events to span multiple epochs, diluting per-epoch predictive clarity. Irrespective of this, combined results underscore that our model can robustly and accurately discriminate between sleep stages while also detecting respiratory and desaturation events. Importantly, these classification tasks are intermediate objectives designed to ensure that the model ‘learns’ a physiologically relevant feature space, forming the basis for our main goal of generating embeddings used towards risk group clustering.

## Novel Sleep Profiles

### Cluster Embeddings into Risk Groups

After training the foundation model, all PSGs passing quality thresholds were transformed into a high-dimensional embedding space that captured its physiological richness and complexity. Clustering analysis, detailed in the Methods section, was performed on these embeddings using k-means, with several permutations of k-values, distance method, and sampling strategy (*Supplementary Figures 1-3*). Silhouette score analysis and consensus matrices indicated that five clusters, based on energy distance and utilizing all samples, represented the largest stable solution (*Figure 2*). The five cluster descriptions were further refined through disease incidence and mortality analyses. Based on clinical outcomes, this process enabled *a posteriori* labeling of risk groups, denoted as RG1 through RG5 (*Figure 2*). Results from Cox regression and propensity score matching, which will be described in subsequent sections, further supported these assignments.

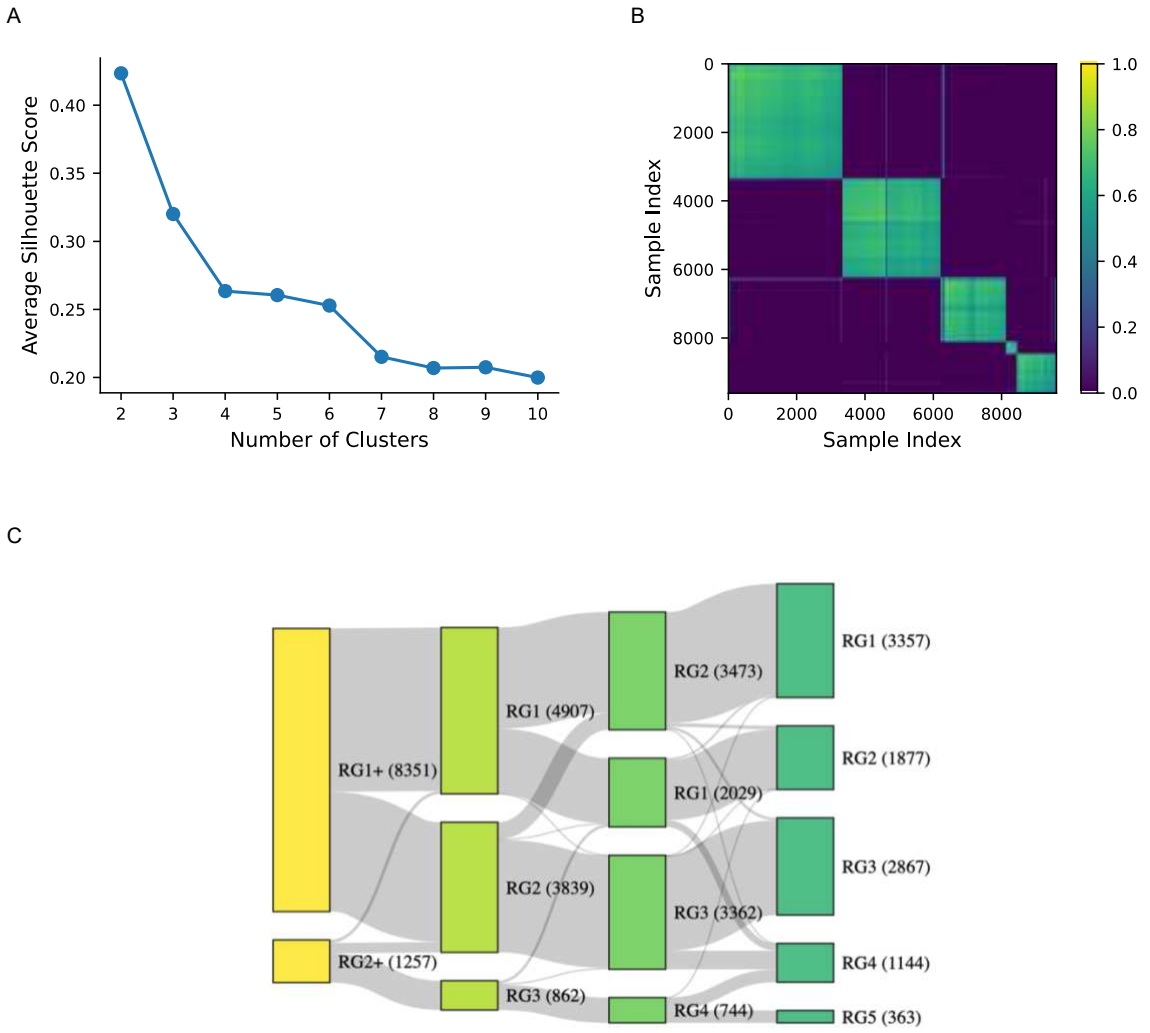


**Figure 2: Quantitative assessment of cluster permutations**. Cluster definitions shown here use projections on all samples and were calculated using energy distance. The average silhouette score for different numbers of clusters is provided **(A)** with the consensus matrix visualization for five clusters shown in **(B)**. The assignment of sleep study samples to each cluster across the 2-, 3-, 4-, and 5-state solutions is shown in a Sankey flow diagram **(C)** where color indicates the number of clusters per permutation.

The five-cluster solution provided the greatest granularity, distinguishing a smaller, high-risk group (RG5) characterized by significant comorbidities and PSG abnormalities consistent with severe sleep disruption. At the other end of the spectrum, RG1 and RG2 represented the lowest-risk groups with minimal abnormalities and comorbidities. Conversely, the more straightforward two-cluster approach isolated a high-risk subgroup, $RG2^{+}$, which largely overlapped with RG5 from the five-cluster solution (*Figure 2C*). Throughout this paper, we focus primarily on the five-cluster solution for its granularity and the uniqueness of its clusters, while also examining $RG2^{+}$ from the two-cluster solution as it corresponds to RG5 in the five-cluster framework.

### Associations between Risk Groups and Incidence of Clinical Outcomes

Cox proportional hazards models adjusting for demographics and relevant comorbidities revealed a strong monotonic trend for disease incidence when comparing clusters from RG1 to RG5 over 14.5±7.1 year observation period (*Figure 3*). Across the five-cluster solution, RG5 consistently exhibited the highest hazard ratios (HR) for cardiovascular and neurologic conditions, including myocardial infarction, MACE, heart failure, atrial fibrillation, mood disorders, epilepsy, and cognitive impairment. Notably, RG3 and RG4 also showed elevated risk compared to the reference group (RG1), although to a lesser extent. The associations persisted after propensity-score matching and trimming for outliers (*Supplementary Table 1*), indicating that the cluster assignments captured meaningful physiological differences not accounted for by simply classifying SDB severity using AHI.

*Figure 3* presents HR and 95% confidence intervals (CIs) for incident clinical outcomes, estimated from Cox regression models adjusted for demographic factors and comorbidities, across five risk groups (RG1 to RG5). RG1 represented the lowest risk and lowest mortality, while RG5 represented the highest risk for the development of adverse clinical outcomes. RG1 was used as the reference group due to its high sample size and low-risk profile, making it a stable and representative baseline for comparison (*Figure 2 and Supplementary Figure 4*). Overall, there was a trend of increasing HRs as risk group number increases, indicating a progressive increase in disease incidence compared to the reference group.


**Figure 3: Hazard ratios of disease incidence among risk groups.** Hazard ratios (HR) were calculated using all available data for each risk group using RG1 as a reference group due to its ordinal ranking from observed lower incidence of mortality and other adverse health outcomes (hence RG1 is not represented in this figure). Fill color is provided for statistically significant HR values such that darker color indicates higher likelihood of disease and worse outcome. The 95% confidence interval values are included parenthetically, and significance levels are indicated with asterisks: *p<0.05; **p<0.01; and ***p<0.001. Abbreviations: MACE: Major Adverse Cardiovascular Events; GERD: Gastroesophageal Reflux Disease.

Most associations in RG2 were close to 1 and not statistically significant, suggesting no substantial increase in disease incidence compared to RG1. Notably, hyperlipidemia (HR=0.75; 95% CI 0.63–0.89; p<0.01) and obesity (HR=0.79; 95% CI 0.68–0.93; p<0.01) in RG2 showed significantly lower incidence of these conditions compared to the reference group.

Patients in RG3 exhibited significant increases in association of incidence of several conditions. There was increased association with diabetes type 2 (HR=1.34; 95% CI 1.12–1.60; p<0.01), hyperlipidemia (HR=1.30; 95% CI 1.08–1.57; p<0.01), major adverse cardiovascular events (MACE) (HR=1.44;

95% CI 1.20–1.73; p<0.001), ischemic heart disease (HR=1.94; 95% CI 1.30–2.88; p<0.001), myocardial infarction (HR=1.42; 95% CI 1.05–1.92; p<0.05), stroke (HR=1.44; 95% CI 1.16–1.77; p<0.001), and mood disorders (HR=1.46; 95% CI 1.24–1.72; p<0.001). These findings suggest that RG3 patients have a moderately increased association of both cardiovascular and neurological conditions compared to RG2.

In RG4, the elevated risks persisted for certain diseases. Significant associations were observed for diabetes type 2 (HR=1.27; 95% CI 1.02–1.58; p<0.05), MACE (HR=1.33; 95% CI 1.06–1.66; p<0.05), coronary artery disease (HR=1.32; 95% CI 1.03–1.70; p<0.05), atrial fibrillation (HR 1.40; 95% CI 1.04–1.88; p<0.05), cognitive impairment (HR=1.42; 95% CI 1.15–1.75; p<0.01), and gastroesophageal reflux disease (GERD) (HR=1.32; 95% CI 1.07–1.64; p<0.05). These results indicate that RG4 patients have a continued elevated risk, particularly for cardiovascular diseases and neurologic outcomes.

The highest risk group, RG5, showed the highest magnitude of association with adverse clinical outcomes. Significant associations were noted for MACE (HR=1.64; 95% CI 1.15–2.32; p<0.01), heart failure (HR=1.65; 95% CI 1.16–2.36; p<0.01), myocardial infarction (HR=1.84; 95% CI 1.16–2.91; p<0.01), atrial fibrillation (HR=2.23; 95% CI 1.47–3.38; p<0.001), cognitive impairment (HR=1.93; 95% CI 1.42–2.62; p<0.001), and epilepsy (HR=2.40; 95% CI 1.46–3.97; p<0.001). These findings highlight a substantially increased risk of serious cardiovascular and neurological conditions in RG5 patients.

To assess the robustness of these findings, a sensitivity analysis was conducted using propensity score matching to control for potential confounding variables (described in Methods Section: Statistical Analyses). *Supplementary Figure 4* summarizes the 6-year disease-free survival rates among propensity score matched patients across all risk groups. Disease free survival rates decreased progressively from RG1 to RG5 for most conditions, reflecting higher incidence rates in higher risk groups. For instance, the 6-year disease-free survival for MACE was 80.6% in RG1, decreasing to 67.3% in RG5 (p<0.001). Similarly, cognitive impairment survival rates declined from 84.3% in RG1 to 75.2% in RG5 (p<0.01).

*Supplementary Table 1* presents HR after propensity score matching. The pattern of increasing association with the clinical outcomes in the higher risk groups remained consistent. Specifically, in RG3, significantly higher HR were observed for hyperlipidemia (HR=1.27; 95% CI 1.05–1.53; p<0.05), diabetes type 2 (HR=1.33; 95% CI 1.11–1.59; p<0.01), MACE (HR=1.35; 95% CI 1.13–1.62; p<0.01), ischemic heart disease (HR=2.01; 95% CI 1.37–2.95; p<0.001), stroke (HR=1.40; 95% CI 1.14–1.72; p<0.01), and mood disorders (HR=1.39; 95% CI 1.18–1.63; p<0.001). In RG4, significant associations were noted for diabetes type 2 (HR=1.38; 95% CI 1.11–1.72; p<0.01), myocardial infarction (HR=1.43; 95% CI 1.03–1.99; p<0.05), atrial fibrillation (HR=1.43; 95% CI 1.04–1.97; p<0.05), and cognitive impairment (HR=1.35; 95% CI 1.08–1.68; p<0.01).

RG5 again demonstrated the highest degree of association with longitudinal clinical outcome development. Significant increased incidence was found for MACE (HR=1.56; 95% CI 1.09–2.24; p<0.05), heart failure (HR=1.53; 95% CI 1.03–2.27; p<0.05), myocardial infarction (HR=1.92; 95% CI 1.19–3.12; p<0.01), atrial fibrillation (HR=1.96; 95% CI 1.24–3.10; p<0.01), cognitive impairment (HR=1.61; 95% CI 1.16–2.24; p<0.01), and epilepsy (HR=1.96; 95% CI 1.16–3.32; p<0.05). These results corroborate the initial findings and confirm that higher risk groups are associated with increased incidence of serious health conditions, even after accounting for confounding factors through propensity score matching. More in depth results for each incident disease are available in *Supplementary Table 2 and 3*.

### All-Cause Mortality

The risk groups also showed clear stratification with respect to all-cause mortality. Kaplan-Meier survival curves (*Figure 4B*) showed that RG5, and to a lesser degree RG4 and RG3, experienced substantially higher mortality rates compared to RG1. This pattern was confirmed with Cox proportional hazards regression (*Table 1*) that adjusted for age, sex, body mass index (BMI), and comorbidities (hypertension, diabetes type 2, heart failure, atrial fibrillation, coronary artery disease, hyperlipidemia). Additional models that excluded those prescribed continuous positive airway pressure (PAP) therapy or patients under 55 years of age demonstrated consistent findings (*Table 1: Models 5 and 6*, respectively). Notably, these results persisted independent of AHI (*Table 1: Model 4*), illustrating that these novel sleep profiles predicted mortality even after accounting for the influence of traditional classification of mild, moderate, or severe SDB.

*Figure 4* presents the survival curves for the risk groups. Panel A shows the Kaplan-Meier curves for the two-cluster solution, highlighting a significant difference in mortality between $RG2^+$ and $RG1^+$ (p=1.2e-42) where $RG2^+$ loosely corresponds to the cohort of RG5 as shown previously (*Figure 2C*). The survival curves for the five risk groups (RG1 to RG5) after propensity score matching are shown in *Figure 4 Panel B*. RG3, RG4, and RG5 exhibit significant differences compared to RG1 (p=2.1e-6, p=3.9e-24, p=1.7e-31), with increasingly divergent survival patterns and a markedly higher risk of all-cause mortality.

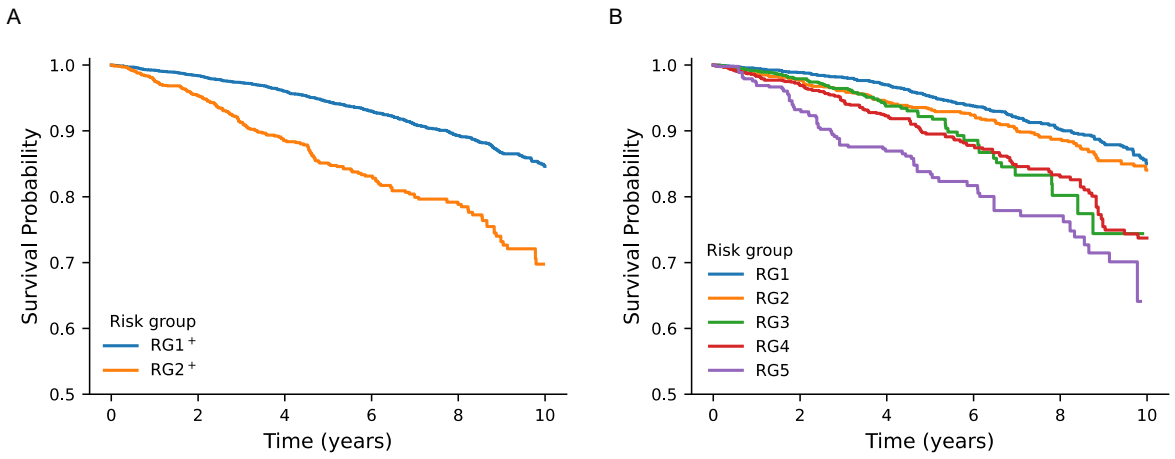



**Figure 4: Kaplan-Meier plot for all-cause mortality of propensity matched cases for different risk groups. (A)** Kaplan-Meier plot for all-cause mortality of propensity-matched cases for the two-cluster solution, $RG1^+$ and $RG2^+$, with p = 1.2e-42. **(B)** Kaplan-Meier plot for all-cause mortality of propensity-matched cases for the five-cluster solution, where risk groups RG3, RG4, and RG5 are significantly different from RG1 (p=2.1e-6, p=3.9e-24, p=1.7e-31). The risk of all-cause mortality for RG5 is significantly higher than other risk groups. The survival probability observed reflects the ordinal ranking of the risk groups.

*Table 1* presents HR for all-cause mortality across risk groups. Six Cox proportional hazards models were employed, each adjusted for different demographic factors and comorbidities, to conduct sensitivity analyses and evaluate the robustness of the observed associations for all-cause mortality. Multiple model permutations were conducted here for all-cause mortality but were not completed for all other assessed comorbidities due to the expansive computational overhead.

Consistent across all models, there was a progressive increase in mortality from RG2 to RG5. In the fully adjusted Model 4 which accounts for age, sex, BMI, comorbidities, and AHI, HRs for all-cause mortality remained significantly elevated: 1.43 for RG2, 1.54 for RG3, 1.75 for RG4, and 2.38 for RG5 (all p-values < 0.01). The persistence of these findings after controlling for AHI demonstrated that the increased mortality associated with higher-risk groups is independent of traditional SDB severity classification. In fact, a Cox regression analysis performed using the same covariates but replacing RG2 to RG5 with mild, moderate and severe sleep apnea defined by AHI thresholds (mild: ≥5 to <15, moderate: ≥15 to <30, severe: ≥30) showed no increase in disease incidence or all-cause mortality with respect to normal AHI (<5) (*Supplementary Table 4*).

These results remained robust in models that excluded patients prescribed PAP treatment (Model 5) and those under 55 years of age (Model 6), suggesting that PAP therapy and older age do not substantially alter the observed associations with increased mortality. The consistent pattern of increasing HR point estimates underscores a strong association between higher-risk group classification and increased all-cause mortality, independent of potential confounders, including AHI.

## Predicting Clusters Using Standard PSG Metrics

We next investigated whether the embedding-based risk groups could be predicted using alternative PSG variables, including commonly used AHI, arousal index, total sleep time, sleep stage percentages, oxygen saturation, and our proposed quantitative definition of "sleep fragmentation." Herein, we define a metric termed sleep fragmentation calculated as the normalized power in the "fast" frequency range of the hypnogram's power spectral density (PSD), specifically transitions occurring faster than 10 minutes. For these variables, a gradient boosting classifier (XGBoost[13]) was trained to predict cluster membership for the two-, three-, four-, and five-cluster solutions using combinations of the top 1 and 5 features ranked by importance. *Supplementary Table 5* summarizes the classification performance and lists the top 5 features for each solution.

For the two-cluster solution, prediction accuracy exceeded 90% using a single feature related to sleep fragmentation. Accuracy further increased and plateaued at 94% when incorporating the top five PSG-derived features. However, for the more complex three-, four-, and five-cluster solutions, prediction accuracy dropped substantially, even when using multiple features. These findings suggest that while simpler risk group structures can be approximated using standard PSG metrics, the full richness of the foundational embeddings and their ability to capture subtle, multidimensional sleep physiology cannot be fully realized by conventional PSG variables alone.

## External Validation

Among the clustering solutions tested, the two-risk group approach demonstrated the highest classification accuracy using standard PSG metrics, with sleep fragmentation alone achieving exceptional predictive performance. Given the strong performance of this simplified two-cluster separation, we applied the sleep fragmentation-based classifier to an external population from the geographically diverse, multicenter Sleep Heart Health Study (SHHS) extracted from the National Sleep Research Resource[8]. Patients assigned to the higher-risk cluster demonstrated a significantly greater likelihood of heart failure and mortality, findings observed in both males ($p = 0.01$ for HF; $p<0.001$ for mortality) and females ($p<0.001$ for HF; $p<0.001$ for mortality) (*Supplementary Figure 6*).

After adjusting for age and sex using Cox proportional hazards regression, $RG2^+$ remained significantly associated with a higher incidence of heart failure (HR=1.52; 95% CI 1.01–2.27; $p < 0.05$) and mortality (HR=1.72; 95% CI 1.34–2.22; $p < 0.001$). Notably, consistent with the baseline characteristics of RG5 in our dataset, patients in $RG2^+$ had markedly low average total sleep time (182.1 minutes, ~3 hours, with a standard deviation of 37.1 minutes). The external validation supports the robustness of the two-cluster risk separation and highlights the clinical relevance of sleep fragmentation as a simple yet meaningful and expedient measure for identifying high-risk patients.

## DISCUSSION

We developed a foundation model that leverages a large clinical PSG registry of multimodal unstructured sleep signals with structured data from EMR, including details of clinical characteristics. This work introduces key technological innovations by employing time-series modeling and clustering techniques to generate unique "sleep embeddings" that can stratify patients and predict clinical outcomes. By capturing the complex interplay between sleep architecture and respiratory event parameters, our approach identifies distinct phenotypic groups associated with adverse clinical outcomes while accounting for confounding factors. To this end, the model was optimized to simultaneously score sleep stages, respiratory events, and desaturations while also achieving results comparable to previous polysomnographic foundation models[14,15]. Our approach leveraged these techniques to reveal distinct risk groups with clinical interpretability and demonstrate clear incremental risk associations. We address a key priority area identified in the American Thoracic Society Workshop[16] that promotes using ML and AI to find novel biomarkers to predict cardiovascular and other health outcomes in sleep disorders.

Among the tested clustering solutions, the two-risk-group approach demonstrated the highest classification accuracy when using standard PSG metrics, particularly solely the sleep fragmentation feature defined herein. This makes the two-cluster solution particularly advantageous for generalization to other datasets, as these metrics, including AHI, sleep stages, and respiratory parameters, are routinely calculated during in-laboratory sleep testing, regardless of the data acquisition system. Notably, one of the clusters, labeled RG2$^+$, remained relatively consistent across different clustering solutions and corresponds closely to RG5 in the five-cluster approach (*Figure 2C*). However, for the three-risk group and higher-order solutions, predictive accuracy decreased substantially (*Supplementary Table 5*). This suggests that more complex risk group solutions based on foundation model embeddings capture nuanced patterns that cannot be easily reproduced using standard PSG metrics alone. Furthermore, when applying the classifier, we developed for the two-cluster solution using standard PSG metrics in the SHHS dataset[17], we observed similar associations of RG2$^+$ from the two-cluster solution with mortality and heart failure, in both sexes. This contrasts with the original SHHS analysis[18], which identified an association only between severe sleep apnea defined by AHI and heart failure in males alone. This external validation underscored the generalizability of RG2$^+$ as a meaningful risk group, suggesting that even basic clustering solutions can be clinically informative while complex solutions provide deeper insights into patient heterogeneity.

This study also highlights the broader limitations of relying solely on AHI as a metric for evaluating SDB presence and severity. While AHI will likely remain an important measure, our findings suggest that additional metrics provide complementary insights into the complexity of SDB. The phenotypic clusters identified in this study align with growing evidence in existing literature that distinct

pathophysiological mechanisms underlying SDB extend beyond what AHI captures in different clinical outcomes; these mechanisms led to outcomes including chronic pain[19], cardiovascular disorders[3], neurodegenerative pathologies[20], and all-cause mortality[21].

A key strength of our study is the use of a large, demographically diverse dataset enriched with minority representation with collection occurring over a decade, containing detailed demographics, and clinical characteristics extracted from the EMR data that was standardized and integrated with PSG information. This comprehensive dataset allowed us to explore the complex relationships between sleep phenotypes and health outcomes with greater precision and granularity. The availability of longitudinal data with mean observation period of 14.5 years further enabled us to investigate long-term associations between the identified risk groups and adverse incident clinical outcomes, such as cardiovascular and neurological diseases. These results have implications for clinical practice and risk stratification and may inform specific treatment approaches. For example, patients in RG5 who demonstrated the highest risk of all-cause mortality may require targeted interventions beyond SDB-targeted therapies. To enhance rigor, we conducted propensity score matching in an attempt to address confounding influences. Furthermore, we conducted external validation of results from this large clinic-based cohort by leveraging a large prospective multicenter population-based cohort, thus supporting reproducibility of the findings.

However, several limitations should be considered. One notable limitation is the lack of objective treatment adherence data which may have influenced clinical outcomes. For example, while PAP prescription information was available, PAP adherence and usage of other therapies for SDB and other sleep disorders were not. Inadequate ascertainment of PAP treatment, however, would be expected to bias findings to the null. Medication usage was not available and may impact the associations discussed. Additionally, while the phenotypic clusters derived from foundation model embeddings demonstrated stronger associations with adverse health outcomes compared to AHI-based classifications, further validation steps are necessary. Further, our findings should be replicated in multiple diverse datasets to ensure their robustness and generalizability. The retrospective nature of the analysis also limits its ability to establish causality, underscoring the need for prospective, case-controlled studies to confirm these results and assess their clinical applicability. Characterizing symptoms and patient reported outcomes with the objective sleep data would be a future area of investigation of clinical relevance and applicability.

In conclusion, our study presents a promising approach to risk stratification in populations undergoing PSG, particularly in the assessment of SDB, leveraging supervised and unsupervised machine learning techniques to extract novel insights from PSG data. The combination of a large dataset and rich patient medical history enabled us to uncover meaningful associations between sleep phenotypes and adverse outcomes. This approach holds promise for improving risk assessment and guiding personalized treatment strategies. Future research should focus on validating these findings

prospectively and exploring the integration of this approach into clinical workflows in implementation science paradigms.

## METHODS

### Data Collection

The STARLIT (Sleep Signals, Testing, and Reports Linked to patient Traits)-10K is a retrospective clinical cohort study approved by the Cleveland Clinic Institutional Review Board (IRB#23-409), assembling a comprehensive database of 10,000 in-laboratory PSG studies from the STARLIT Registry[22], along with patients' corresponding EMR data. The sleep studies were conducted at the Cleveland Clinic between January 2012 and December 2022. For each visit, Nihon Kohden equipment and Polysmith software were used to collect sleep physiological data, sleep staging information, epoch-by-epoch annotations, and the sleep study report finalized by a board-certified sleep physician. The first 1,000 PSGs were selected randomly, and the remaining 9,000 PSGs were intentionally selected to reflect a broad range of ages, temporal distribution, and enriched minority representation over the entire cohort. Clinical characteristic data were extracted from EMRs using natural language processing from various sources, including hospital admissions, patient encounters, problem lists, referrals, and surgical logs. All-cause mortality was determined by integrating multiple sources, including the EMRs for the most current data, the Ohio Death Index, and the Social Security Death Index. The median duration of medical history available for each patient was 9.4 (IQR: 4.6, 13.6) years before PSG and 4.4 (IQR: 2.0, 5.7) years after testing, for a total of 15.1 (IQR: 9.4, 19.0) years for a mean observation period of 14.5 ± 7.1 years. An overall data description, including the multi-institutional collaboration that uses STARLIT-10K is described in Bilal et al.[23]

### Data Preprocessing

Each PSG contains data from sensors measuring multiple physiological aspects of sleep, including EEG, EOG, EMG, EKG, respiratory airflow, thoracic and abdominal movement, $EtCO_2$, snoring, and $SpO_2$. The PSG data were preprocessed to remove artifact and ensure consistency across studies. The time-series data were resampled to 128 Hz and preprocessed as described in Brink-Kjaer et al.[24] Signals were filtered using infinite impulse response (IIR) filters to eliminate artifacts and ensure that signals contained similar spectral content across recordings. The IIR filters were implemented as elliptic filters with an order of 16, a maximum passband ripple of 1 dB, and a minimum stopband attenuation of 40 dB. The cut-off frequencies for the filters were: EEG and EOG: band-pass (0.3–45 Hz); EMG: high-pass (10 Hz); EKG: high-pass (0.3 Hz); nasal pressure: high-pass (0.1 Hz); airflow and plethysmography belts: band-pass (0.1–15 Hz); and blood oxygen saturation: no filtering. All filters were applied forwards and backwards to avoid signal phase distortion. Finally, the signal amplitudes, except blood oxygen saturation, were normalized such that −1 and 1 corresponded to the 5th and 95th percentiles. The blood oxygen saturation was normalized to −1 and 1 corresponding to 60% and 100% saturation.

### Model Training and Embedding Generation

The time-series foundation model was trained to address three primary tasks: sleep stage classification, respiratory event detection which included both apneas and hypopneas, and desaturation event identification. A total of 9,203 PSGs were used for training, with 30 samples forming a validation set and 500 samples set aside as a final test set. The details of PSG inclusion criteria are provided in *Supplementary Methods: Data Quality Assessment*.

The model was optimized using the Adam optimizer[25] with a learning rate of 1e-5 and trained for 50 epochs with a batch size of 500 samples. Loss functions were task-specific, using cross-entropy loss for sleep stage classification and binary cross-entropy loss for desaturation and respiratory event detection. The final loss was computed as the average of the task-specific losses to ensure balanced optimization across tasks. The model's internal representation size was set to 768 which is the default representation dimension from the RoBERTa-base model, with 12 transformer layers contributing to a total of 126 million trainable parameters.

Final sleep embeddings were generated by stacking sequential 30-second epoch representations of the CLS tokens from each sleep study into a final two-dimensional embedding matrix.

### Clustering Method

After training the model, the learned CLS token served as an embedding for each 30-second segment of sleep, creating a two-dimensional representation of the PSG data with 768 rows and columns corresponding to the number of 30-second epochs in each recording. Consequently, the embedding dimensions varied between samples due to differences in PSG duration. These differences precluded the direct application of k-means clustering. To address this and compress the feature space, each sample was projected onto all other samples by calculating pairwise distances between its embedding and the embeddings of all other samples. Distances were computed for corresponding rows of the embeddings (representing the time dimension) and averaged across the embedding dimension to produce a single distance value for each pair. This process generated a feature vector for each sample, where each element represented its distance to another sample. These feature vectors were then used as input for k-means clustering.

Given the variability in PSG duration, distance metrics capable of comparing distributions with differing numbers of samples were required. Metrics such as energy distance[26] and Earth mover's distance,[27] both of which are well-suited for comparing distributions of varying sizes, were evaluated. Additionally, the analysis included projections restricted to just the test samples to assess the effect of limiting the sample space.

Clustering performance was assessed using silhouette scores and consensus matrices. The silhouette score quantified the cohesion and separation of clusters, with higher scores indicating better-defined groups. The consensus matrix was generated through subsampling, where 80% of samples and

features were randomly selected in 100 iterations. For each pair of samples, the matrix recorded the proportion of iterations in which the samples were assigned to the same cluster. To facilitate interpretation, the consensus matrix was ordered using hierarchical clustering with the Ward criterion,[28] grouping samples based on their co-clustering frequency to visually highlight stable cluster structures.

*Supplementary Figure 1* presents the silhouette scores and consensus heatmaps for clustering based on energy distance applied to all samples, the version selected for subsequent analyses in the study. The heatmaps reveal a clear block-diagonal structure up to five clusters, indicating stable and well-defined groupings. Beyond five clusters, this structure deteriorates, suggesting reduced clustering quality.

A sensitivity analysis was conducted by varying the distance metric and projection size. *Supplementary Figure 2* presents the analysis using energy distance with projections limited to test samples, while *Supplementary Figure 3* shows the results for Earth mover's distance applied to test sample projections.

*Supplementary Figure 5* illustrates the sample assignment across the three 5-cluster solutions, highlighting the robustness of the five-cluster solution to variations in distance metric and projection size.

These results demonstrate that using energy distance across all samples provides robust clustering, as evidenced by the stable block-diagonal consensus structure for up to five clusters and consistently high silhouette scores. This clustering configuration was subsequently used for further analyses throughout the study.

### Statistical Analyses

Baseline characteristics were compared between risk groups using chi-square tests for categorical variables, with Bonferroni corrections for pairwise comparisons on significant results ($p < 0.05$). Continuous characteristics were analyzed using Welch's ANOVA, with significant differences further examined via pairwise *t*-tests with Bonferroni correction. Logistic regression was used to assess associations between total sleep time on the PSG and baseline medical history, adjusting for age, BMI, sex, and relevant comorbidities (*Supplementary Table 6*).

Survival time was defined as the duration from the date of baseline PSG until the date of death or, for censored cases, the date of last follow-up as recorded in the EMR. Cox proportional hazards regression was used to examine the association between demographic variables (age, sex, BMI), comorbidities, AHI, and clustered risk groups with time-to-event outcomes for various diseases. Relevant comorbidities (*Supplementary Table 6*) were included for each disease, and only patients with at least 5 years of prior medical history were considered, excluding those with prior occurrences of the disease. The proportional odds assumption was met for all models.

A sensitivity analysis was conducted by performing a second Cox regression on weighted propensity score-matched data to evaluate whether the results of the initial analysis were influenced by residual confounding or imbalances in baseline characteristics. Propensity scores were estimated using multinomial logistic regression to control for confounding across risk groups using age, BMI, sex, relevant comorbidities (*Supplementary Table 6*), and number of years of medical history available before sleep testing. Propensity weights were calculated as stabilized inverse propensity scores[29] and were trimmed at the 5th and 95th percentiles to minimize bias introduced by extreme values. Kaplan-Meier analysis estimated disease-free survival probabilities at 2, 4, and 6 years, with log-rank tests comparing survival rates between the reference and other risk groups.


### Acknowledgments

This project was supported by the Cleveland Clinic-IBM Discovery Accelerator program. Matheus Araujo and Reena Mehra are partially supported by Grant Number 1R21HL170206-01 from the National Heart Lung and Blood Institute. The Sleep Heart Health Study (SHHS) was supported by National Heart, Lung, and Blood Institute cooperative agreements U01HL53916 (University of California, Davis), U01HL53931 (New York University), U01HL53934 (University of Minnesota), U01HL53937 and U01HL64360 (Johns Hopkins University), U01HL53938 (University of Arizona), U01HL53940 (University of Washington), U01HL53941 (Boston University), and U01HL63463 (Case Western Reserve University). The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002).


## References


1. Berry, R.B.*, et al.* Rules for Scoring Respiratory Events in Sleep: Update of the 2007 AASM Manual for the Scoring of Sleep and Associated Events. *Journal of Clinical Sleep Medicine* **08**, 597-619 (2012).
2. Zinchuk, A.V.*, et al.* Polysomnographic phenotypes and their cardiovascular implications in obstructive sleep apnoea. *Thorax* **73**, 472-480 (2018).
3. Mazzotti, D.R.*, et al.* Symptom Subtypes of Obstructive Sleep Apnea Predict Incidence of Cardiovascular Outcomes. *Am J Respir Crit Care Med* **200**, 493-506 (2019).
4. Miller, C.B.*, et al.* Clusters of Insomnia Disorder: An Exploratory Cluster Analysis of Objective Sleep Parameters Reveals Differences in Neurocognitive Functioning, Quantitative EEG, and Heart Rate Variability. *Sleep* **39**, 1993-2004 (2016).
5. Ahlqvist, E.*, et al.* Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *Lancet Diabetes Endocrinol* **6**, 361-369 (2018).

6. Young, A.L.*, et al.* Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. *Nat Commun* **9**, 4273 (2018).

7. Vaswani, A. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).

8. Quan, S.F.*, et al.* The Sleep Heart Health Study: Design, Rationale, and Methods. *Sleep* **20**, 1077-1085 (1997).

9. Zhang, G.-Q.*, et al.* The National Sleep Research Resource: towards a sleep data commons. *Journal of the American Medical Informatics Association* **25**, 1351-1358 (2018).

10. Zhou, T., Niu, P., Sun, L. & Jin, R. One fits all: Power general time series analysis by pretrained lm. *Advances in neural information processing systems* **36**, 43322-43355 (2023).

11. Liu, Y. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* **364**(2019).

12. Vallat, R. & Walker, M.P. An open-source, high-performance tool for automated sleep staging. *Elife* **10**(2021).

13. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, San Francisco, California, USA, 2016).

14. Thapa, R.*, et al.* Sleepfm: Multi-modal representation learning for sleep across ecg, eeg and respiratory signals. in *AAAI 2024 Spring Symposium on Clinical Foundation Models* (2024).

15. Fox, B.*, et al.* A foundational transformer leveraging full night, multichannel sleep study data accurately classifies sleep stages. *Sleep*, zsaf061 (2025).

16. Cohen, O.*, et al.* The Great Controversy of Obstructive Sleep Apnea Treatment for Cardiovascular Risk Benefit: Advancing the Science Through Expert Consensus. An Official American Thoracic Society Workshop Report. *Ann Am Thorac Soc* **22**, 1-22 (2024).

17. Quan, S.F.*, et al.* The Sleep Heart Health Study: design, rationale, and methods. *Sleep* **20**, 1077-1085 (1997).

18. Gottlieb, D.J.*, et al.* Prospective study of obstructive sleep apnea and incident coronary heart disease and heart failure: the sleep heart health study. *Circulation* **122**, 352-360 (2010).

19. Charokopos, A., Card, M.E., Gunderson, C., Steffens, C. & Bastian, L.A. The Association of Obstructive Sleep Apnea and Pain Outcomes in Adults: A Systematic Review. *Pain Medicine* **19**, S69-S75 (2018).

20. Yeo, B.S.Y.*, et al.* The association of obstructive sleep apnea with blood and cerebrospinal fluid biomarkers of Alzheimer's Dementia-A systematic review and meta-analysis. *Sleep Medicine Reviews* **70**, 101790 (2023).

21. Lin, Y., *et al.* Objective Sleep Duration and All-Cause Mortality Among People With Obstructive Sleep Apnea. *JAMA Network Open* **6**, e2346085-e2346085 (2023).

22. Heinzinger, C.M., *et al.* Sleep-Disordered Breathing, Hypoxia, and Pulmonary Physiologic Influences in Atrial Fibrillation. *J Am Heart Assoc* **12**, e031462 (2023).

23. Bilal, E., *et al.* A novel approach for phenotypic characterization of sleep disorders. in *2024 IEEE International Conference on Digital Health (ICDH)* 8-13 (2024).

24. Brink-Kjaer, A., *et al.* Age estimation from sleep studies using deep learning predicts life expectancy. *npj Digital Medicine* **5**, 103 (2022).

25. Kingma, D.P. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

26. Székely, G. E-statistics: energy of statistical samples (Tech. Rep. No. 02-16). *Bowling Green State University, Dep. Math. stat* (2002).

27. Rubner, Y., Tomasi, C. & Guibas, L.J. A metric for distributions with applications to image databases. in *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)* 59-66 (IEEE, 1998).

28. Ward Jr, J.H. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* **58**, 236-244 (1963).

29. Xu, S., *et al.* Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals. *Value Health* **13**, 273-277 (2010).

30. Araujo, M., *et al.* Machine Learning Electroencephalography Biomarkers Predictive of Epworth Sleepiness Scale. (Cold Spring Harbor Laboratory, 2022).

31. Levitt, J., *et al.* Automated detection of electroencephalography artifacts in human, rodent and canine subjects using machine learning. *Journal of Neuroscience Methods* **307**, 53-59 (2018).

32. Azarbarzin, A., *et al.* The hypoxic burden of sleep apnoea predicts cardiovascular disease-related mortality: the Osteoporotic Fractures in Men Study and the Sleep Heart Health Study. *Eur Heart J* **40**, 1149-1157 (2019).

33. Azarbarzin, A., *et al.* The Sleep Apnea-Specific Pulse-Rate Response Predicts Cardiovascular Morbidity and Mortality. *Am J Respir Crit Care Med* **203**, 1546-1555 (2021).

34. Kwon, Y., *et al.* Lung to finger circulation time in sleep study and coronary artery calcification: the Multi-Ethnic Study of Atherosclerosis. *Sleep Med* **75**, 8-11 (2020).

35. Powell-Wiley, T.M., *et al.* Obesity and Cardiovascular Disease: A Scientific Statement From the American Heart Association. *Circulation* **143**, e984-e1010 (2021).

36. Jia, G. & Sowers, J.R. Hypertension in Diabetes: An Update of Basic Mechanisms and Clinical Disease. *Hypertension* **78**, 1197-1205 (2021).

37. Aune, D.*, et al.* Blood pressure, hypertension and the risk of atrial fibrillation: a systematic review and meta-analysis of cohort studies. *Eur J Epidemiol* **38**, 145-178 (2023).

38. Nakanishi, R.*, et al.* Relationship of Hypertension to Coronary Atherosclerosis and Cardiac Events in Patients With Coronary Computed Tomographic Angiography. *Hypertension* **70**, 293-299 (2017).

39. Chruściel, P.*, et al.* Associations between the lipid profile and the development of hypertension in young individuals - the preliminary study. *Arch Med Sci* **18**, 25-35 (2022).

40. Feingold, K.R. Dyslipidemia in diabetes. (2015).

41. Alonso, A.*, et al.* Blood Lipids and the Incidence of Atrial Fibrillation: The Multi-Ethnic Study of Atherosclerosis and the Framingham Heart Study. *Journal of the American Heart Association* **3**, e001211.

42. Yao, Y.S., Li, T.D. & Zeng, Z.H. Mechanisms underlying direct actions of hyperlipidemia on myocardium: an updated review. *Lipids Health Dis* **19**, 23 (2020).

43. Alloubani, A., Nimer, R. & Samara, R. Relationship between Hyperlipidemia, Cardiovascular Disease and Stroke: A Systematic Review. *Curr Cardiol Rev* **17**, e051121189015 (2021).

44. Oh, G.C. & Cho, H.-J. Blood pressure and heart failure. *Clinical Hypertension* **26**, 1 (2020).

45. Elendu, C.*, et al.* Heart failure and diabetes: Understanding the bidirectional relationship. *Medicine (Baltimore)* **102**, e34906 (2023).

46. Wang, T.J.*, et al.* Temporal relations of atrial fibrillation and congestive heart failure and their joint influence on mortality: the Framingham Heart Study. *Circulation* **107**, 2920-2925 (2003).

47. Rush, C.J.*, et al.* Prevalence of Coronary Artery Disease and Coronary Microvascular Dysfunction in Patients With Heart Failure With Preserved Ejection Fraction. *JAMA Cardiol* **6**, 1130-1143 (2021).

48. Echouffo-Tcheugui, J.B.*, et al.* Care Patterns and Outcomes in Atrial Fibrillation Patients With and Without Diabetes: ORBIT-AF Registry. *J Am Coll Cardiol* **70**, 1325-1335 (2017).

49. Yan, T.*, et al.* Coronary Artery Disease and Atrial Fibrillation: A Bidirectional Mendelian Randomization Study. *J Cardiovasc Dev Dis* **9**(2022).

50. Vgontzas, A., Cui, L. & Merikangas, K.R. Are sleep difficulties associated with migraine attributable to anxiety and depression? *Headache* **48**, 1451-1459 (2008).

51. Ojeda, B., *et al.* Understanding the different relationships between mood and sleep disorders in several groups of non-oncological patients with chronic pain. *Curr Med Res Opin* **34**, 669-676 (2018).

52. Pearson, O., *et al.* The relationship between sleep disturbance and cognitive impairment in mood disorders: A systematic review. *J Affect Disord* **327**, 207-216 (2023).

53. Perini, G., *et al.* Cognitive impairment in depression: recent advances and novel treatments. *Neuropsychiatr Dis Treat* **15**, 1249-1258 (2019).

54. Banjade, P., *et al.* The Interplay between Obstructive Sleep Apnea, Chronic Obstructive Pulmonary Disease, and Congestive Heart Failure: Time to Collectively Refer to Them as Triple Overlap Syndrome? *Medicina (Kaunas)* **59**(2023).

55. McCracken, L.M. & Iverson, G.L. Disrupted sleep patterns and daily functioning in patients with chronic pain. *Pain Res Manag* **7**, 75-79 (2002).

56. Guerrera, C.S., *et al.* A narrative review on insomnia and hypersomnolence within Major Depressive Disorder and bipolar disorder: A proposal for a novel psychometric protocol. *Neurosci Biobehav Rev* **158**, 105575 (2024).

## SUPPLEMENTARY RESULTS

### Baseline analysis of novel risk groups

*Supplementary Table 7* presents the demographic and clinical characteristics, along with the sleep parameters defined in *Supplementary Table 8*, for patients across five risk groups, RG1 to RG5. The data are organized under several subheadings: Sociodemographics, Cardiovascular Risk Factors, Cardiovascular Disease, Neurological Disorders, Other Medical History, PSG results, Alternative Metrics, and Other measures.

Risk Group 1 (RG1) had the largest sample size (n=3,357) with a mean age of 50.8 years. This group had a relatively balanced sex distribution, with 52.6% males. The prevalence of cardiovascular risk factors such as hypertension (59.8%), hyperlipidemia (53.5%), and type 2 diabetes (32.7%) was moderate compared to other groups. PSG results indicated mild SDB, with an average AHI of 12.4 events/hour and an arousal index of 23.8 events/hour. This group also showed moderate levels relative to other RGs of sleep fragmentation and hypoxic burden.

Risk Group 2 (RG2) consisted of younger patients with a mean age of 44.0 years and the lowest percentage of males (37.6%). This group exhibited the lowest prevalence of cardiovascular risk factors: hypertension (47.5%), hyperlipidemia (37.8%), and type 2 diabetes (24.7%). PSG findings revealed the mildest SDB of all groups, with AHI of 5.4 events/hour, and the lowest arousal index

(19.3 events/hour). Neurological disorders such as migraine (19.7%) and mood disorders (50.9%) were relatively more prevalent in this group.

Risk Group 3 (RG3) had patients with a mean age of 49.7 years and an intermediate sex distribution (48.7% males). This group showed higher prevalence of mood disorders (56.6%) and chronic pain (44.8%). Cardiovascular risk factors were similar to RG1, with hypertension present in 59.7% of patients. PSG results indicated mild SDB with an AHI of 11.2 events/hour and an arousal index of 22.7 events/hour.

Risk Group 4 (RG4) included older patients with a mean age of 58.6 years and the highest percentage of males among larger groups (60.8%). This group exhibited the highest prevalence of cardiovascular risk factors: hypertension (75.6%), hyperlipidemia (65.2%), and type 2 diabetes (41.8%). Cardiovascular diseases were more common, with heart failure present in 18.4% and atrial fibrillation in 15.3% of patients. PSG results showed increased SDB severity relative to lower risk groups, with an AHI of 22.7 events/hour and a significantly higher arousal index of 41.5 events/hour, suggesting more fragmented sleep.

Similar to RG4, Risk Group 5 (RG5), the smallest group (n=363), had patients with high mean age (58.6 years) and the highest proportion of males (66.4%). This group demonstrated the most severe PSG abnormalities, with the highest AHI (37.3 events/hour) and arousal index (60.8 events/hour), suggesting severe SDB and sleep fragmentation. Cardiovascular diseases were most prevalent in this group, with heart failure in 21.8% and atrial fibrillation in 17.9% of patients. Despite severe PSG findings, the prevalence of some neurological disorders, such as mood disorders (51.0%), was similar to other groups.

To further evaluate the associations between risk groups and various diseases, *Supplementary Tables 9* and *10* provide the odds ratios of comorbidities for patients in risk groups RG2 to RG5 relative to RG1, derived from logistic regression analyses adjusted for age, BMI, and relevant comorbidities (as outlined in *Supplementary Table 6*). These findings largely corroborate the observations from the baseline table, with the exception of RG3, which was significantly associated with increased odds for all the diseases examined. This may be attributed to a strong age effect, given the relatively younger age of individuals in this group.

In summary, the risk groups exhibited a gradient of increasing age, male predominance, cardiovascular risk factors, and severity of SDB from RG2 to RG5. RG2 was characterized by younger age, a lower prevalence of cardiovascular risks, and milder PSG findings. RG4 and RG5 included older patients with higher cardiovascular morbidity and more severe PSG abnormalities, reflecting a higher risk profile. RG1 and RG3 shared similar PSG characteristics; however, RG3 distinguished itself by being associated with a significantly higher number of comorbidities after adjusting for demographic factors.

## SUPPLEMENTARY METHODS

### Data Quality Assessment

Sample records involving study type that was not manually confirmed as PSG were excluded. This exclusion encompassed split studies, home sleep apnea tests, multiple sleep latency tests, maintenance of wakefulness tests, and other non-PSG studies in which sleep was intentionally or artificially altered.

Data quality assessment for the 10,000 files was performed using data from six referenced channels. In an ideal setup, odd-numbered channels, positioned on the left side of the scalp, were referenced by 'M1', while even-numbered channels, positioned on the right side, were referenced by 'M2'. In rare cases of reference issues, the reference from the opposite side was employed. The electroencephalography (EEG) channels utilized were 'F3', 'F4', 'C3', 'C4', 'O1', and 'O2'[30].

A high-pass filter with a passband frequency of 1 Hz and a notch filter with a stop-band of 57.5–62.5 Hz was applied to all recordings using a zero-phase forward and reverse digital filter. Waveforms in each channel were divided into 1-second epochs, and each epoch was checked for artifacts using a previously validated machine learning method by Levitt et al. [31]. Three separate artifact detection scans were applied to each file, corresponding to the three non-rapid eye movement sleep stages: N1, N2, and N3. Rapid eye movement (REM) sleep was excluded due to the presence of ocular artifacts. Each scan started at the first occurrence of the respective sleep stage in the file and continued for a duration of two minutes.

In each sleep stage, the percentage of clean files was determined based on the artifact detection algorithm's results. A file was labeled as clean if it had at least four clean channels out of the six EEG channels, with each channel considered clean if it had at least 60% artifact-free data. The percentages of clean files for N1, N2, and N3 sleep stages were 90%, 88%, and 91%, respectively. A total of 266 files that were not artifact-free in all three sleep stages were excluded.

### Polysomnogram Features

The aggregated PSG metrics were extracted from the structured HTML format of the final sleep study report submitted by the provider to the EMR. The metrics parsed from the original file include total and REM AHI, arousal index, obstructive apneas, central apneas, hypopneas, mean oxygen saturation, minimum and maximum SpO2, sleep time with oxygen saturation less than 90% $SpO_2$, maximum $EtCO_2$, percentage of sleep time in each sleep stage, total sleep time, total REM time, and snoring. All conventional PSG variables were computed after scoring by Nihon Kohden's Polysmith 12 software. To enrich the analyses, the following alternative metrics were computed: sleep apnea-specific hypoxic burden,[32] sleep apnea-specific pulse-rate response (Delta HR),[33] lung to finger circulation time,[34] and sleep fragmentation, which is defined herein and calculated as the normalized

power in the "fast" frequency range of the hypnogram's power spectral density (PSD), specifically transitions occurring faster than 10 minutes.

Providers are instructed to remove text variables reporting zero values in the final PSG report if the value was not measurable. This may result in missing values that should be zero. With regard to data cleanliness and pre-processing, missing values were imputed where appropriate to address clinical data incompleteness. The following corrections were applied: sleep time with oxygen saturation $<O_2$ under 90% was set to 0 if the minimum oxygen saturation was above 90%. If the central apnea index (CAI) was zero, the number of central apneas was set to zero. To ensure accuracy, the reported AHI total was validated by recomputing the sum of individual counts of respiratory events (obstructive apneas, central apneas, mixed apneas, and hypopneas) based on total sleep time. In case of missing respiratory events and AHI total was zero, individual respiratory events counts were set to zero. When total sleep time was zero, the AHI total was set to N/A, as no sleep was observed.

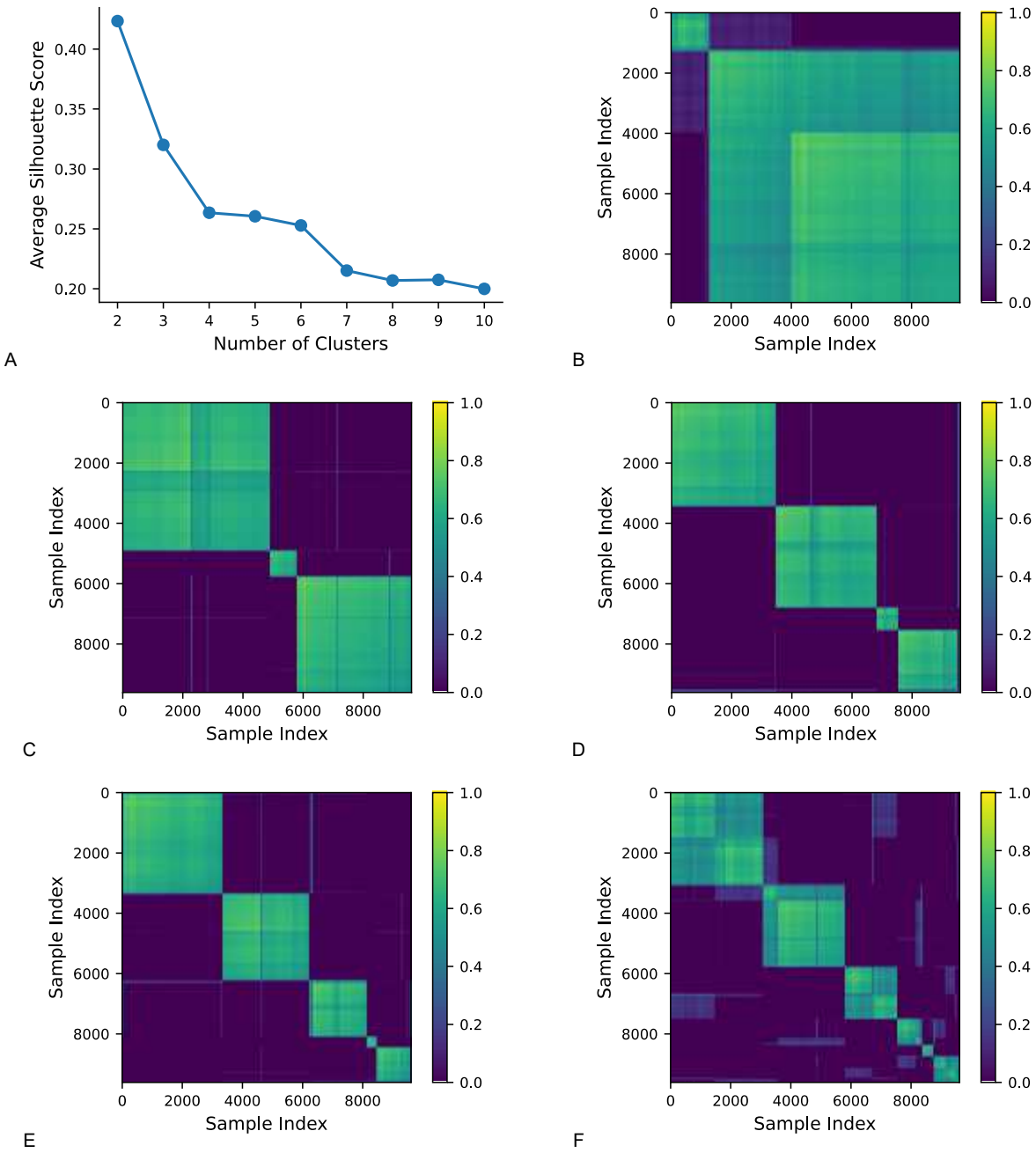
### Disease Vocabulary and Comorbidities

To define the disease vocabulary used in our analyses, we curated a comprehensive list of clinically significant conditions with their respective diagnosis code (ICD-10) that are relevant to SDB, sleep architecture, and associated health risks (*Supplementary Table 11)*. This vocabulary included disease outcomes: type 2 diabetes, hypertension, hyperlipidemia, heart failure (HF), atrial fibrillation (AF), migraine, mood disorders, cognitive impairment, chronic obstructive pulmonary disease (COPD), chronic pain, gastroesophageal reflux disease (GERD), chronic insomnia, and major adverse cardiovascular events (MACE) which included HF, myocardial infarction, coronary artery disease (CAD), coronary artery bypass grafting, and stroke. Mood disorders included depression, anxiety, bipolar disorder, post-traumatic stress disorder, and related conditions. Cognitive impairment included cognitive deficits, amnesia, and dementia-related conditions such as Alzheimer's, Parkinson's, Lewy body disease, and vascular dementia. Chronic pain included chronic pain syndrome and fibromyalgia. These conditions were identified from diagnoses recorded in the Electronic Health Record (Epic®) that appeared at least once in our cohort. They were selected based on both established and hypothesized relationships with sleep physiology and their relevance to long-term health outcomes, with the goal of providing a broad view of how sleep-related risk factors may influence diverse health domains.

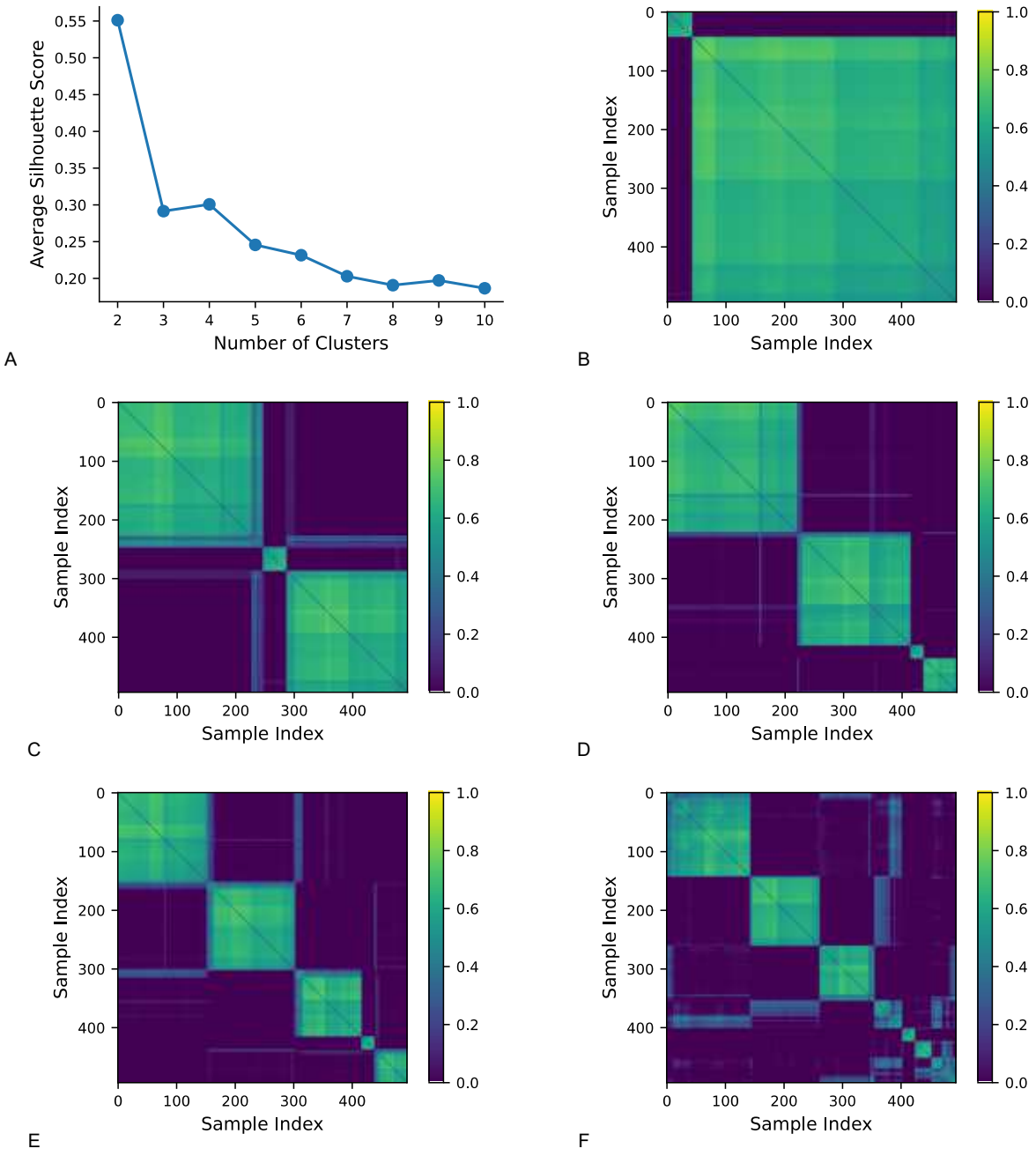
For each disease outcome, we individually selected covariates to adjust for relevant confounders that could potentially affect the association between the risk groups and outcomes (Supplementary Table 12). All models included obesity as a covariate, highlighting the importance this condition has in cardiovascular and neurologic diseases[35]. Covariate selection was based upon biological plausibility of factors considered to be confounding influences in the respective models. **Type 2 diabetes** was modeled with hypertension as a covariate, reflecting the well-known role this factor plays in the development and progression of diabetes[36]. **Hypertension** was modeled with diabetes, AF,[37] HF ,

CAD,[38] and hyperlipidemia, [39] acknowledging the multifactorial nature of hypertension and its associations with both metabolic and cardiovascular conditions. **Hyperlipidemia** was modeled with hypertension,[39] diabetes,[40] AF,[41] HF, [42] and CAD,[43] recognizing the intertwined relationships between lipid metabolism and these cardiovascular risk factors. **HF** was modeled with hypertension,[44] diabetes,[45] AF,[46] CAD,[47] and hyperlipidemia,[42] as these factors collectively contribute to the development and worsening of HF. **AF** was modeled with hypertension,[37] diabetes,[48] HF,[46] CAD,[49] and hyperlipidemia,[41] given the strong associations between these conditions and the risk of arrhythmias. **Migraine** was modeled with mood disorders, reflecting the interplay between these factors in the pathophysiology of migraine and sleep disorders.[50] **Mood disorders** were modeled with chronic pain, recognizing how these conditions are often co-occurring and can exacerbate each other's impact on mental and physical health.[51] **Cognitive impairment** was modeled with mood disorders, as this factor may be implicated in the progression of cognitive decline.[52,53] **COPD** was modeled with HF, as obesity and HF can be major contributors to the development and severity of COPD.[54] **Chronic pain** was modeled with mood disorders, emphasizing how chronic pain often coexists with these conditions, influencing the overall health burden.[51] **GERD** was modeled with hypertension and diabetes, reflecting the well-documented associations between these conditions and the development of reflux. **Chronic insomnia** was modeled with chronic pain[55] and mood disorders,[56] highlighting complex interactions that contribute to the persistence of insomnia.[55,56] **MACE** was modeled without obesity and with hypertension, diabetes, AF, and hyperlipidemia, to best capture the risk factors pertinent to these serious outcomes. This tailored approach ensured that each model accounts for disease-specific confounders, providing a robust and nuanced framework for assessing the relationship between the identified risk groups and a wide range of clinically meaningful health outcomes.

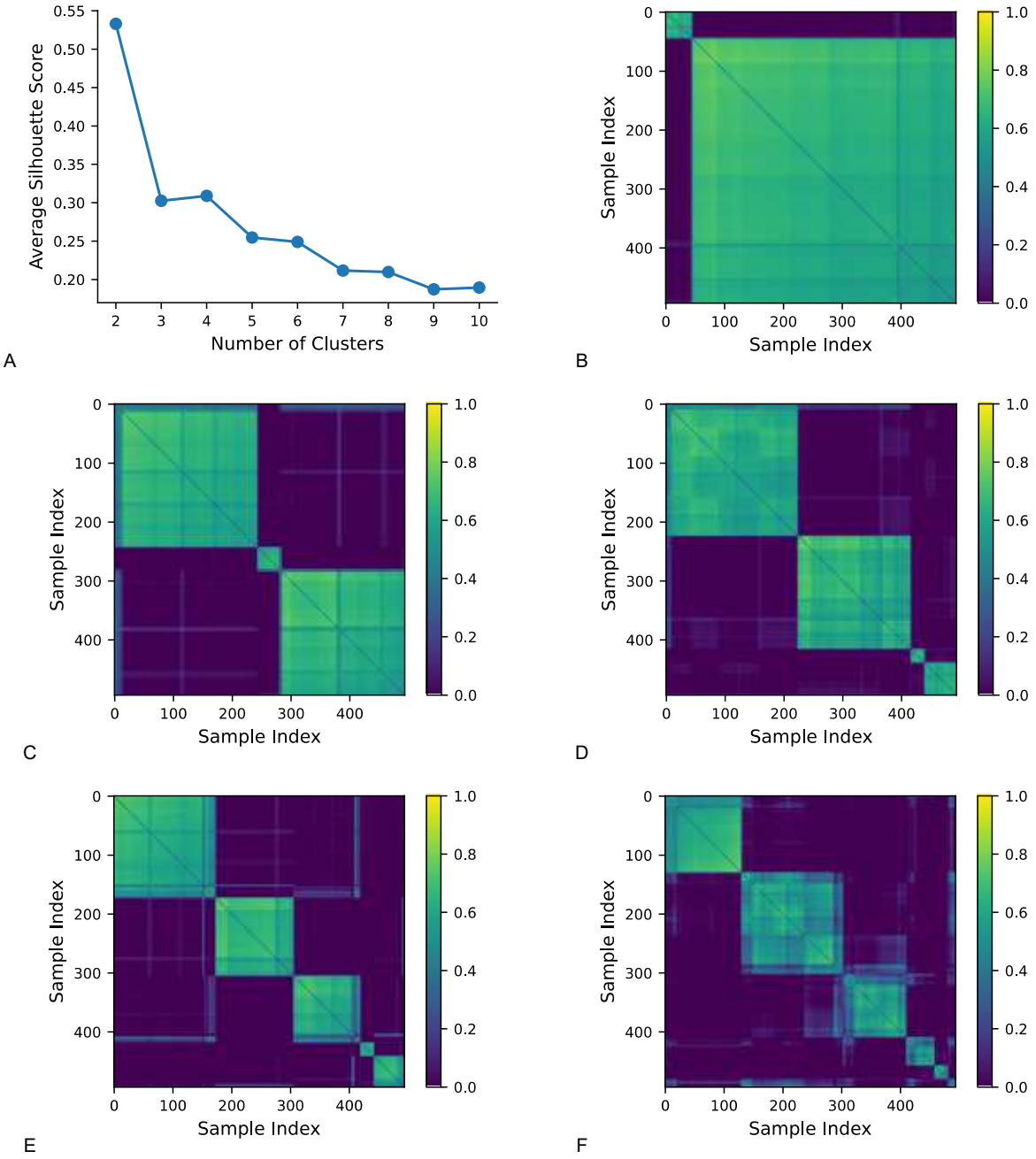
## Supplementary Figures



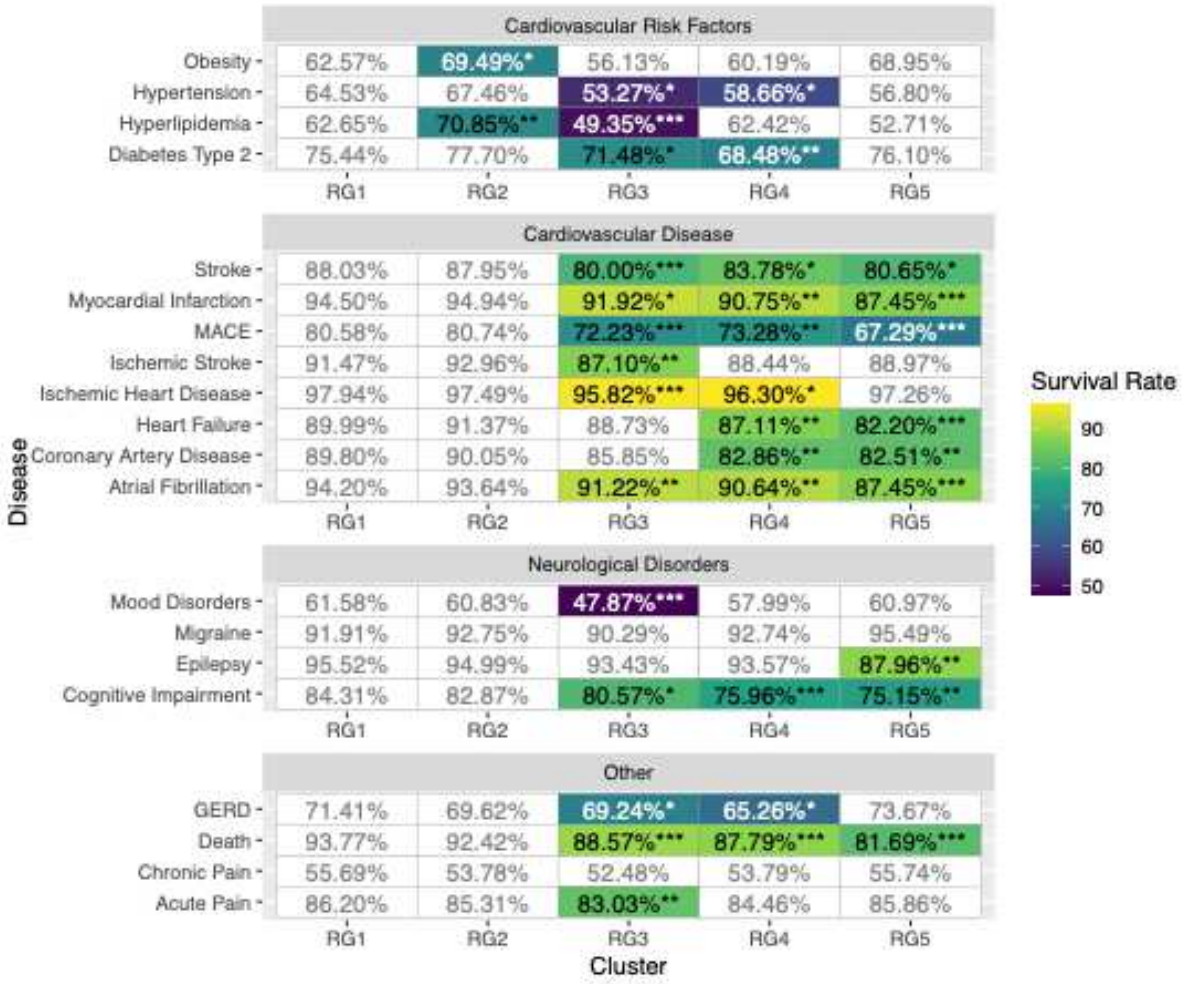

**Supplementary Figure 1: Sensitivity analyses of different numbers of clusters derived using energy distance and all samples projection.** Clusters use projections on all samples using the energy distance. (A) Silhouette score for different number of clusters, consensus matrices visualization for 2 (B), 3 (C), 4 (D), 5 (E) and 6 (F) clusters. Note panel A and E are reproduced here from Figure 2 to show all k-values and enable a side-by-side comparison.



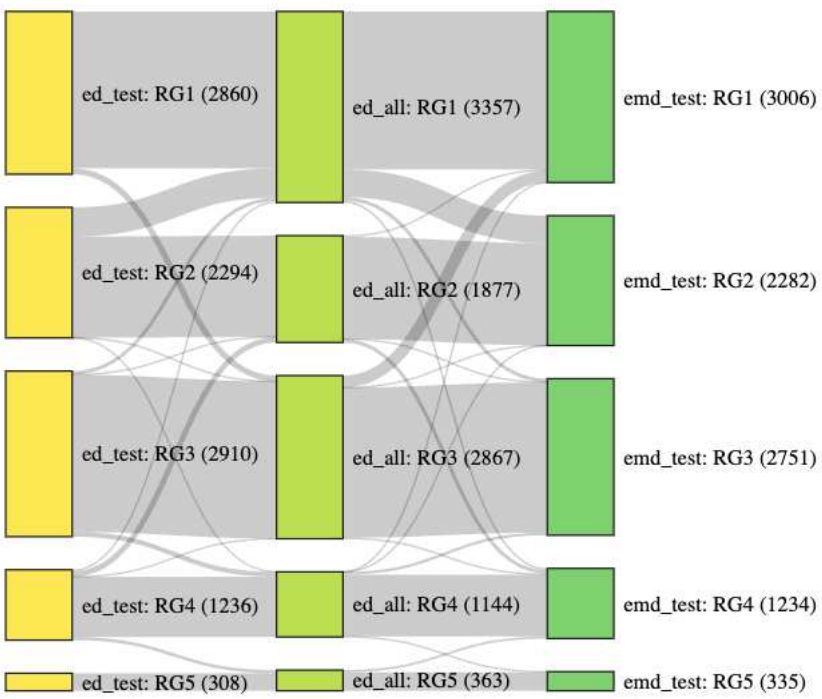

**Supplementary Figure 2: Sensitivity analyses of different numbers of clusters derived using energy distance and test samples projection.** Clusters use projections on test samples using the energy distance. (A) Silhouette score for different number of clusters, consensus matrices visualization for 2 (B), 3 (C), 4 (D), 5 (E) and 6 (F) clusters



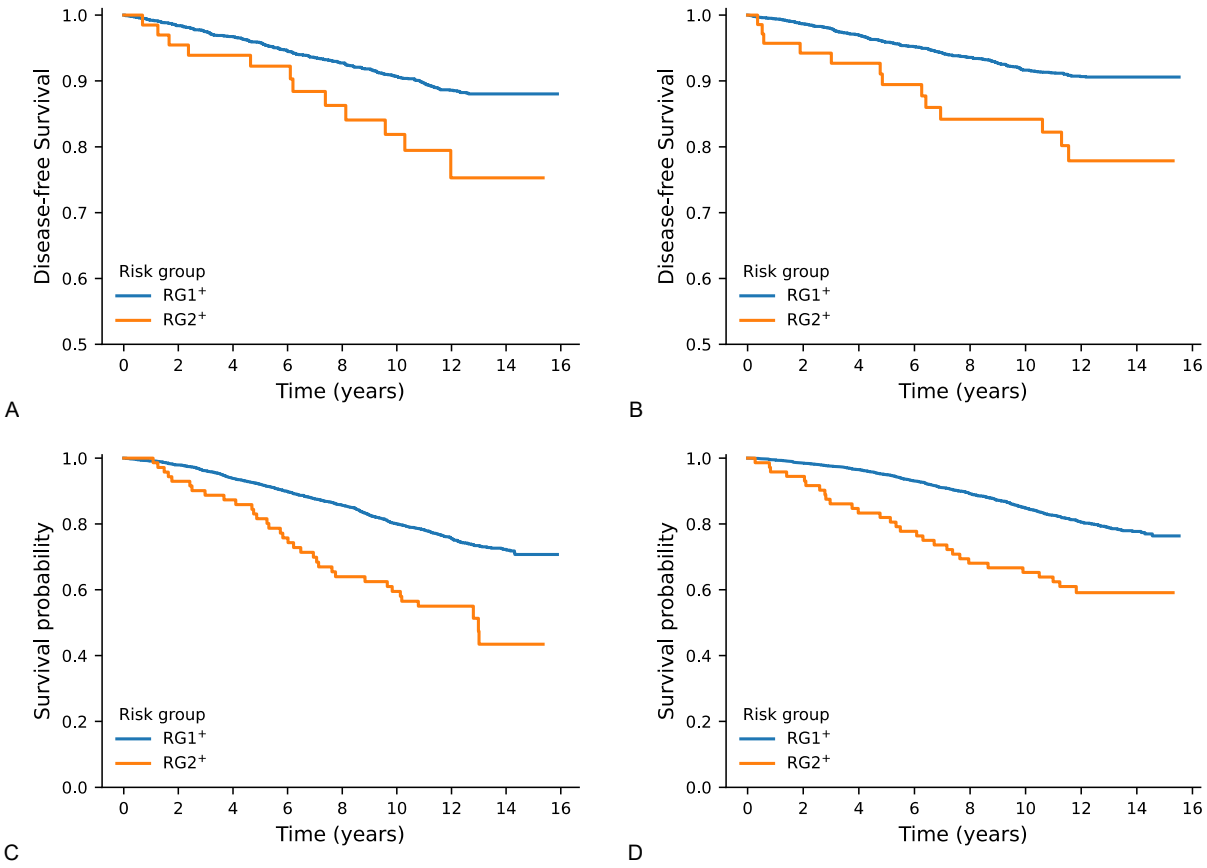

**Supplementary Figure 3: Sensitivity analyses of different numbers of clusters derived using Earth mover's distance and test samples projection.** Clusters use projections on test samples using the Earth mover's distance. (A) Silhouette score for different numbers of clusters, consensus matrices visualization for (B) two, (C) three, (D) four, (E) five, and (F) six clusters.




**Supplementary Figure 4: Summary of 6-year disease free survival among propensity score matched patients from different risk groups.** Survival rate was calculated after six years from the date of baseline PSG. Fill color is provided for statistically significant survival rates with weighted propensity score matching such that darker color indicates worse outcome (lower survival due to disease indicated on y-axis). Significance levels are indicated with asterisks: *p<0.05; **p<0.01; and ***p<0.001. Abbreviations: MACE: Major Adverse Cardiovascular Events; GERD: Gastroesophageal Reflux Disease.




**Supplementary Figure 5: Sankey diagram for the sample assignment between different five cluster solutions.** The three columns represent the different five-cluster solutions from *Supplementary figures 1 – 3*



**Supplementary Figure 6: Kaplan-Meier plots of risk groups RG1$^+$ and RG2$^+$ from the two-cluster solution applied to the Sleep Heart Health Study data.** Congestive heart failure-free survival plots are shown in panel A for males (p = 0.0103) and panel B for females (p = 0.0008) for risk groups RG1$^+$ and RG2$^+$ of the two-cluster solution applied to the data from the Sleep Heart Health Study. Survival plots for all-cause mortality are shown in panels C for males (p = 3.8e-07) and panel D for females (p = 1.7e-06).

## Tables

Table 1: Hazard ratios of all-cause mortality among patients from different risk groups

| Model | No samples | No events | RG2 | RG3 | RG4 | RG5 |
|---|---|---|---|---|---|---|
| **1** | 7043 | 544 | 0.92 (0.72-1.17) | 1.69 (1.29-2.23)*** | 3.11 (2.46-3.92)*** | 4.83 (3.60-6.50)*** |
| **2** | 7043 | 544 | 1.54 (1.20-1.98)*** | 1.65 (1.25-2.16)*** | 1.84 (1.45-2.33)*** | 2.73 (2.02-3.69)*** |
| **3** | 7043 | 544 | 1.47 (1.14-1.88)** | 1.52 (1.16-2.00)** | 1.67 (1.31-2.11)*** | 2.16 (1.59-2.95)*** |
| **4** | 7043 | 544 | 1.43 (1.11-1.84)** | 1.54 (1.17-2.03)** | 1.75 (1.37-2.23)*** | 2.38 (1.73-3.28)*** |
| **5** | 3700 | 299 | 1.32 (0.95-1.82) | 1.66 (1.14-2.40)** | 1.78 (1.28-2.47)*** | 1.76 (1.12-2.78)* |
| **6** | 1569 | 523 | 1.53 (1.15-2.04)** | 1.47 (1.08-2.01)* | 1.67 (1.30-2.15)*** | 1.75 (1.24-2.49)** |

Model 1: without adjustment; 2: adjusted for age, sex and body mass index (BMI); 3: adjusted for age, sex, BMI and comorbidities (hypertension, diabetes type 2, heart failure, atrial fibrillation, coronary artery disease, hyperlipidemia); 4: adjusted for age, sex, BMI, comorbidities and apnea hypopnea index (AHI); 5: adjusted for age, sex, BMI, comorbidities and excluding patients for which PAP therapy was prescribed; 6: adjusted for age, sex, BMI and comorbidities excluding patients less than 55 years of age.

## Supplementary Tables

Supplementary Table 1: Hazard ratios of incident diseases and all-cause mortality among patients from different risk groups after propensity score matching

| Disease | RG2 | RG3 | RG4 | RG5 |
|---|---|---|---|---|
| **Cardiovascular Risk Factors** | | | | |
| **Hypertension** | 1.08 (0.90-1.30) | 1.16 (0.94-1.43) | 1.26 (0.94-1.69) | 1.31 (0.79-2.16) |
| **Hyperlipidemia** | 0.74 (0.63-0.88)*** | 1.27 (1.05-1.53)* | 1.02 (0.79-1.32) | 1.12 (0.77-1.63) |
| **Diabetes Type 2** | 0.93 (0.79-1.10) | 1.33 (1.11-1.59)** | 1.38 (1.11-1.72)** | 1.05 (0.70-1.57) |
| **Obesity** | 0.79 (0.67-0.92)** | 1.10 (0.93-1.30) | 0.82 (0.66-1.02) | 0.90 (0.61-1.33) |
| **Cardiovascular Disease** | | | | |
| **MACE** | 1.12 (0.95-1.33) | 1.35 (1.13-1.62)*** | 1.23 (0.96-1.57) | 1.56 (1.09-2.24)* |
| **Heart Failure** | 1.03 (0.82-1.29) | 1.07 (0.83-1.36) | 1.23 (0.93-1.62) | 1.53 (1.03-2.27)* |
| **Myocardial Infarction** | 1.17 (0.88-1.55) | 1.42 (1.06-1.90)* | 1.43 (1.03-1.99)* | 1.92 (1.19-3.12)** |
| **Ischemic Heart Disease** | 1.05 (0.68-1.61) | 2.01 (1.37-2.95)*** | 1.40 (0.86-2.30) | 1.28 (0.56-2.97) |
| **Coronary Artery Disease** | 1.03 (0.83-1.28) | 1.15 (0.91-1.44) | 1.29 (0.98-1.70) | 1.44 (0.93-2.22) |
| **Atrial Fibrillation** | 1.12 (0.85-1.47) | 1.35 (1.02-1.80)* | 1.43 (1.04-1.97)* | 1.96 (1.24-3.10)** |
| **Stroke** | 1.12 (0.92-1.35) | 1.40 (1.14-1.72)** | 0.97 (0.75-1.27) | 1.31 (0.87-1.95) |
| **Ischemic Stroke** | 0.95 (0.75-1.21) | 1.32 (1.03-1.68)* | 1.01 (0.74-1.38) | 1.02 (0.61-1.70) |
| **Neurological Disorders** | | | | |
| **Mood Disorders** | 1.03 (0.88-1.21) | 1.39 (1.18-1.63)*** | 1.16 (0.95-1.43) | 1.24 (0.89-1.73) |
| **Migraine** | 0.81 (0.62-1.06) | 1.16 (0.88-1.53) | 1.06 (0.75-1.51) | 0.75 (0.40-1.40) |
| **Cognitive Impairment** | 1.13 (0.95-1.35) | 1.17 (0.96-1.42) | 1.35 (1.08-1.68)** | 1.61 (1.16-2.24)** |
| **Epilepsy** | 1.07 (0.78-1.47) | 1.09 (0.77-1.56) | 1.28 (0.86-1.92) | 1.96 (1.16-3.32)* |
| **Other** | | | | |
| **GERD** | 1.05 (0.89-1.25) | 1.20 (1.00-1.43)* | 1.36 (1.10-1.69)** | 0.92 (0.62-1.37) |
| **Chronic Pain** | 1.03 (0.91-1.16) | 1.06 (0.92-1.22) | 0.99 (0.84-1.18) | 0.98 (0.74-1.29) |

| | | | | |
|---|---|---|---|---|
| **Acute Pain** | 1.06 (0.88-1.27) | 1.36 (1.12-1.66)** | 1.14 (0.89-1.46) | 1.08 (0.71-1.62) |
| **Death** | 1.49 (1.19-1.87)*** | 1.54 (1.19-2.00)** | 1.67 (1.28-2.19)*** | 2.49 (1.77-3.50)*** |

Significance levels: *p<0.05; **p<0.01; ***p<0.001

Supplementary Table 2: Cox regression analysis results for each disease

See attached file: cox_summary_results.xlsx

Supplementary Table 3: Propensity score matching analysis for each disease

See attached file: propensity_analysis_results.xlsx

Supplementary Table 4: Hazard ratios of incident diseases and all-cause mortality among patients from different AHI categories

| Disease | AHI Mild | AHI Moderate | AHI Severe |
|---|---|---|---|
| **Cardiovascular Risk Factors** | | | |
| **Hypertension** | 0.89 (0.75-1.06) | 0.91 (0.72-1.14) | 1.00 (0.75-1.32) |
| **Diabetes Type 2** | 0.97 (0.83-1.13) | 0.90 (0.74-1.09) | 1.19 (0.96-1.49) |
| **Hyperlipidemia** | 1.26 (1.07-1.48)** | 1.32 (1.07-1.62)** | 1.29 (1.01-1.66)* |
| **Obesity** | 1.21 (1.04-1.41)* | 0.94 (0.77-1.14) | 1.31 (1.06-1.63)* |
| **Cardiovascular Disease** | | | |
| **MACE** | 0.84 (0.72-0.99)* | 0.74 (0.61-0.90)** | 0.76 (0.61-0.97)* |
| **Heart Failure** | 0.83 (0.67-1.02) | 0.65 (0.50-0.84)** | 0.80 (0.60-1.07) |
| **Myocardial Infarction** | 0.96 (0.74-1.25) | 0.79 (0.58-1.09) | 1.20 (0.85-1.69) |
| **Ischemic Heart Disease** | 1.31 (0.90-1.91) | 0.91 (0.58-1.45) | 0.98 (0.58-1.66) |
| **Coronary Artery Disease** | 1.02 (0.83-1.26) | 0.91 (0.71-1.16) | 1.00 (0.75-1.33) |
| **Atrial Fibrillation** | 0.99 (0.77-1.27) | 0.75 (0.55-1.02) | 0.76 (0.53-1.10) |
| **Stroke** | 0.78 (0.65-0.93)** | 0.79 (0.63-0.97)* | 0.75 (0.57-0.97)* |
| **Ischemic Stroke** | 0.87 (0.70-1.08) | 0.78 (0.60-1.02) | 0.68 (0.49-0.96)* |
| **Neurological Disorders** | | | |
| **Mood Disorders** | 0.95 (0.82-1.10) | 0.83 (0.70-0.99)* | 0.79 (0.64-0.98)* |
| **Migraine** | 1.18 (0.93-1.49) | 1.07 (0.78-1.46) | 0.92 (0.61-1.41) |
| **Cognitive Impairment** | 0.87 (0.74-1.02) | 0.71 (0.58-0.87)*** | 0.70 (0.54-0.89)** |
| **Epilepsy** | 0.89 (0.67-1.17) | 0.65 (0.45-0.95)* | 0.59 (0.36-0.95)* |
| **Other** | | | |
| **GERD** | 0.95 (0.81-1.10) | 0.81 (0.67-0.99)* | 0.73 (0.58-0.93)* |
| **Chronic Pain** | 0.93 (0.83-1.04) | 0.86 (0.74-0.99)* | 0.79 (0.66-0.95)* |
| **Acute Pain** | 0.92 (0.78-1.09) | 0.91 (0.74-1.13) | 0.73 (0.55-0.97)* |
| **Death** | 0.71 (0.58-0.87)** | 0.52 (0.40-0.68)*** | 0.77 (0.58-1.02) |

Disease abbreviations: AHI: apnea hypopnea index, MACE: Major Adverse Cardiovascular Events; GERD: Gastroesophageal Reflux Disease

Significance levels: *p<0.05; **p<0.01; ***p<0.001

Values are Hazard Ratio (95% Confidence Interval)

Supplementary Table 5: Classification of risk groups using standard sleep measures

| Model accuracy | 2-cluster solution | 3-cluster solution | 4-cluster solution | 5-cluster solution |
|---|---|---|---|---|
| **XGB w/ top-1 features** | 93.0% | 54.6% | 41.6% | 41.7% |
| **XGB w/ top-5 features** | 94.1% | 67.8% | 56.3% | 56.2 |
| **Top-5 features** | Sleep fragmentation<br>Total sleep time<br>AHI total<br>Total REM time<br>Total NREM time | Sleep fragmentation<br>Total sleep time<br>Hypoxic burden<br>Total N3 time<br>AHI total | Sleep fragmentation<br>Total sleep time<br>AHI total<br>Hypoxic burden<br>Mean oxygen saturation | Sleep fragmentation<br>Total sleep time<br>AHI total<br>Hypoxic burden<br>Mean oxygen saturation |

The top-5 features are in order of their importance.

Abbreviations: AHI: apnea hypopnea index, XGB: XGBoost, N3: non-rapid eye movement sleep stage 3, NREM: non-rapid eye movement sleep stage, REM: rapid eye movement sleep stage

Supplementary Table 6: Comorbidities used as covariates for each disease in logistic and cox regression analyses

See attached file: disease_comorbidities.csv

Supplementary Table 7: Demographic and clinical characteristics of patients from different risk groups

| | Risk group – No. (%) or Mean (Std) | | | | | |
|---|---|---|---|---|---|---|
| **Characteristic** | RG**1 (n=3357)** | RG**2 (n=1877)** | RG**3 (n=2867)** | RG**4 (n=1144)** | RG**5 (n=363)** | **P-value** |
| **Sociodemographics** | | | | | | |
| **Age (years)** | 50.8 (14.9)[4] | 44.0 (15.0)[4] | 49.7 (15.8)[4] | 58.6 (16.1)[3] | 58.6 (16.8)[3] | **<.001** |
| **Caucasian** | 1581 (47.1%) | 765 (40.8%) | 1231 (42.9%) | 575 (50.3%) | 159 (43.8%) | |
| **African-American** | 1140 (34.0%) | 697 (37.1%) | 952 (33.2%) | 395 (34.5%) | 154 (42.4%) | |
| **Asian** | 162 (4.8%) | 104 (5.5%) | 170 (5.9%) | 40 (3.5%) | 16 (4.4%) | |
| **Multiracial** | 429 (12.8%) | 276 (14.7%) | 469 (16.4%) | 119 (10.4%) | 32 (8.8%) | |
| **BMI (kg/m²)** | 33.9 (8.2)[1] | 32.7 (9.1)[4] | 33.8 (8.7)[1] | 34.4 (9.6)[1] | 35.4 (10.5)[1] | **<.001** |
| **Male** | 1766 (52.6%) | 705 (37.6%) | 1397 (48.7%) | 696 (60.8%) | 241 (66.4%) | |
| **Height (cm)** | 170.7 (10.6)[2] | 168.7 (10.5)[4] | 169.9 (10.9)[4] | 171.0 (11.0)[2] | 171.7 (11.1)[2] | **<.001** |
| **Weight (kg)** | 98.8 (24.9)[2] | 92.9 (26.0)[4] | 97.3 (25.9)[3] | 100.8 (29.9)[2] | 104.5 (32.3)[3] | **<.001** |
| **Neck Circum. (cm)** | 40.1 (4.6)[3] | 38.4 (4.4)[4] | 39.9 (4.7)[3] | 41.2 (5.4)[3] | 42.1 (5.7)[3] | **<.001** |
| **Cardiovascular Risk Factors** | | | | | | |
| **Diabetes Type 1** | 89 (2.7%)[2] | 48 (2.6%)[2] | 96 (3.3%)[2] | 61 (5.3%)[3] | 26 (7.2%)[3] | **<.001** |
| **Diabetes Type 2** | 1099 (32.7%)[4] | 463 (24.7%)[4] | 1080 (37.7%)[2] | 478 (41.8%)[2] | 162 (44.6%)[2] | **<.001** |
| **Hypertension** | 2008 (59.8%)[3] | 892 (47.5%)[4] | 1712 (59.7%)[3] | 865 (75.6%)[3] | 281 (77.4%)[3] | **<.001** |
| **Hyperlipidemia** | 1796 (53.5%)[2] | 710 (37.8%)[4] | 1500 (52.3%)[2] | 746 (65.2%)[3] | 216 (59.5%)[1] | **<.001** |
| **Obesity** | 1490 (44.4%)[4] | 681 (36.3%)[4] | 1592 (55.5%)[2] | 592 (51.7%)[2] | 197 (54.3%)[2] | **<.001** |
| **Cardiovascular Disease** | | | | | | |
| **Heart Failure** | 250 (7.4%)[3] | 122 (6.5%)[3] | 297 (10.4%)[4] | 210 (18.4%)[3] | 79 (21.8%)[3] | **<.001** |
| **Myocardial Infarction** | 142 (4.2%)[3] | 59 (3.1%)[3] | 190 (6.6%)[4] | 111 (9.7%)[3] | 44 (12.1%)[3] | **<.001** |
| **Atrial Fibrillation** | 249 (7.4%)[3] | 90 (4.8%)[4] | 264 (9.2%)[3] | 175 (15.3%)[3] | 65 (17.9%)[3] | **<.001** |
| **Stroke** | 319 (9.5%)[3] | 137 (7.3%)[3] | 395 (13.8%)[3] | 206 (18.0%)[3] | 70 (19.3%)[2] | **<.001** |
| **Ischemic Stroke** | 229 (6.8%)[3] | 98 (5.2%)[3] | 280 (9.8%)[4] | 152 (13.3%)[3] | 54 (14.9%)[3] | **<.001** |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Coronary Artery Disease** | 450 (13.4%)[3] | 175 (9.3%)[4] | 421 (14.7%)[3] | 301 (26.3%)[3] | 95 (26.2%)[3] | **<.001** |
| **Ischemic Heart Disease** | 45 (1.3%)[3] | 12 (0.6%)[3] | 68 (2.4%)[3] | 38 (3.3%)[2] | 18 (5.0%)[3] | **<.001** |
| **Acute Coronary Disease** | 2 (0.1%) | 2 (0.1%) | 4 (0.1%) | 3 (0.3%) | 0 (0.0%) | 0.503 |
| **MACE** | 784 (23.4%)[4] | 333 (17.7%)[4] | 789 (27.5%)[4] | 499 (43.6%)[3] | 152 (41.9%)[3] | **<.001** |
| **Neurological Disorders** | | | | | | |
| **Migraine** | 486 (14.5%)[2] | 369 (19.7%)[3] | 565 (19.7%)[3] | 131 (11.5%)[2] | 39 (10.7%)[2] | **<.001** |
| **Alzheimer's Disease** | 8 (0.2%)[1] | 4 (0.2%)[1] | 9 (0.3%)[1] | 19 (1.7%)[3] | 4 (1.1%) | **<.001** |
| **Mood Disorders** | 1526 (45.5%)[3] | 956 (50.9%)[2] | 1624 (56.6%)[2] | 610 (53.3%)[1] | 185 (51.0%) | **<.001** |
| **Cognitive Impairment** | 475 (14.1%)[3] | 284 (15.1%)[2] | 626 (21.8%)[2] | 259 (22.6%)[2] | 76 (20.9%)[1] | **<.001** |
| **Epilepsy** | 172 (5.1%)[1] | 129 (6.9%) | 242 (8.4%)[1] | 84 (7.3%) | 24 (6.6%) | **<.001** |
| **Other Medical History** | | | | | | |
| **COPD** | 375 (11.2%)[2] | 178 (9.5%)[3] | 360 (12.6%)[2] | 218 (19.1%)[3] | 65 (17.9%)[2] | **<.001** |
| **Chronic Pain** | 883 (26.3%)[3] | 477 (25.4%)[3] | 1284 (44.8%)[4] | 404 (35.3%)[3] | 121 (33.3%)[3] | **<.001** |
| **GERD** | 1264 (37.7%)[2] | 646 (34.4%)[2] | 1256 (43.8%)[2] | 486 (42.5%)[2] | 141 (38.8%) | **<.001** |
| **CABG** | 69 (2.1%)[2] | 30 (1.6%)[2] | 74 (2.6%)[2] | 63 (5.5%)[3] | 24 (6.6%)[3] | **<.001** |
| **Chronic Insomnia** | 315 (9.4%)[3] | 184 (9.8%)[3] | 547 (19.1%)[2] | 195 (17.0%)[2] | 64 (17.6%)[2] | **<.001** |
| **PSG results** | | | | | | |
| **AHI Total** | 12.4 (9.9)[4] | 5.4 (6.4)[4] | 11.2 (12.3)[4] | 22.7 (24.4)[4] | 37.3 (39.0)[4] | **<.001** |
| **AHI Off-supine** | 8.3 (9.5)[3] | 3.5 (7.0)[4] | 7.9 (11.2)[3] | 18.1 (23.6)[4] | 35.6 (40.4)[4] | **<.001** |
| **AHI Supine** | 21.3 (23.0)[4] | 10.2 (16.5)[4] | 18.3 (22.2)[4] | 31.6 (32.6)[4] | 41.4 (44.4)[4] | **<.001** |
| **Arousal Index** | 23.8 (12.2)[4] | 19.3 (11.1)[4] | 22.7 (13.2)[4] | 41.5 (24.4)[4] | 60.8 (42.0)[4] | **<.001** |
| **CAI** | 0.3 (1.2)[4] | 0.1 (0.4)[4] | 0.5 (1.9)[4] | 1.4 (5.6)[3] | 3.3 (12.2)[3] | **<.001** |
| **Central Apneas** | 1.7 (6.2)[4] | 0.9 (4.8)[4] | 2.5 (9.9)[3] | 5.1 (23.9)[3] | 5.5 (24.3)[2] | **<.001** |
| **Hypopneas** | 57.8 (48.9)[3] | 27.2 (32.5)[3] | 48.8 (56.7)[3] | 57.8 (76.7)[3] | 37.8 (84.4)[2] | **<.001** |
| **Maximum $EtCO_2$** | 48.7 (10.3)[1] | 49.4 (20.0)[1] | 49.0 (9.0)[2] | 48.1 (6.5)[2] | 46.4 (8.3)[4] | **<.001** |
| **Mean $O_2$ Saturation** | 94.0 (2.1)[3] | 95.4 (1.6)[4] | 94.2 (2.2)[3] | 93.6 (2.5)[3] | 93.4 (3.0)[3] | **<.001** |
| **Min $O_2$ Saturation** | 84.9 (6.2)[2] | 89.0 (4.9)[4] | 85.0 (7.6)[2] | 84.1 (7.6)[3] | 83.8 (10.7)[1] | **<.001** |
| **NREM AHI** | 9.5 (9.6)[3] | 3.7 (5.8)[4] | 9.2 (11.9)[3] | 20.9 (24.6)[4] | 37.1 (39.7)[4] | **<.001** |
| **Obstructive Apneas** | 7.7 (15.9)[3] | 2.2 (6.6)[4] | 7.4 (17.2)[3] | 16.8 (38.5)[3] | 21.2 (59.9)[3] | **<.001** |
| **REM AHI** | 24.9 (21.2)[4] | 11.9 (14.5)[4] | 22.3 (21.8)[4] | 30.9 (29.0)[3] | 36.3 (36.8)[3] | **<.001** |
| **Sleep Stage N1 (%)** | 8.1 (5.6)[4] | 7.0 (5.6)[4] | 8.6 (6.2)[4] | 16.7 (12.6)[4] | 33.7 (27.1)[4] | **<.001** |
| **Sleep Stage N2 (%)** | 68.3 (10.3)[4] | 66.5 (12.1)[4] | 71.1 (9.9)[3] | 70.5 (14.3)[3] | 57.5 (26.8)[4] | **<.001** |
| **Sleep Stage N3 (%)** | 6.8 (8.1)[4] | 10.1 (10.4)[4] | 4.7 (7.3)[4] | 3.9 (7.9)[3] | 2.5 (8.1)[3] | **<.001** |
| **Sleep Stage REM (%)** | 17.4 (7.5)[3] | 17.1 (8.1)[3] | 15.9 (7.6)[4] | 9.4 (9.0)[4] | 3.9 (8.1)[4] | **<.001** |
| **Sleep Time $SpO_2$ Under 90% (%)** | 7.1 (19.8)[4] | 2.0 (10.2)[4] | 11.7 (33.6)[2] | 14.9 (33.8)[2] | 14.4 (33.2)[2] | **<.001** |
| **Snoring** | 3076 (91.8%)[2] | 1588 (84.9%)[4] | 2563 (90.1%)[2] | 1013 (89.3%)[2] | 276 (76.5%)[4] | **<.001** |
| **Total Sleep Time (min)** | 328.7 (61.1)[4] | 342.6 (71.6)[4] | 321.4 (59.9)[4] | 201.3 (73.9)[4] | 98.4 (89.0)[4] | **<.001** |
| **Total NREM time (min)** | 270.3 (48.8)[3] | 282.6 (57.2)[4] | 269.4 (50.2)[3] | 181.7 (67.5)[4] | 91.8 (78.6)[4] | **<.001** |
| **Total REM time (min)** | 58.6 (29.8)[3] | 60.2 (34.1)[3] | 52.2 (28.5)[4] | 19.7 (21.2)[4] | 6.9 (16.5)[4] | **<.001** |
| **Alternative metrics** | | | | | | |
| **Sleep Fragmentation** | 0.2 (0.1)[3] | 0.2 (0.1)[4] | 0.2 (0.1)[3] | 0.4 (0.1)[4] | 0.7 (0.2)[4] | **<.001** |
| **Hypoxic Burden** | 24.8 (22.1)[4] | 8.5 (10.1)[4] | 27.5 (32.6)[4] | 55.9 (69.2)[4] | 128.7 (164.6)[4] | **<.001** |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Lung 2 finger time** | 20.5 (6.3)[2] | 20.0 (7.7)[2] | 20.2 (7.0)[2] | 22.5 (8.2)[3] | 23.8 (10.4)[3] | **<.001** |
| **Delta HR** | 12.8 (5.6) | 12.8 (5.9) | 13.1 (5.9) | 12.6 (7.5) | 13.8 (8.6) | 0.062 |
| **Other** | | | | | | |
| **ESS** | 8.7 (5.3)[3] | 9.4 (5.6)[4] | 8.7 (5.3)[3] | 7.7 (5.2)[3] | 7.9 (5.4)[3] | **<.001** |
| **Average Sleep[+] (hours)** | 6.2 (2.1) | 6.4 (1.9) | 6.5 (2.0) | 6.3 (2.4) | 5.9 (2.2) | 0.189 |

Abbreviations: AHI: Apnea Hypopnea Index; BMI: Body Mass Index; CABG: Coronary Artery Bypass Grafting; CAI: Central Apnea Index; COPD: Chronic Obstructive Pulmonary Disease; ESS: Epworth Sleepiness Scale; $EtCO_2$: End-Tidal Carbon Dioxide; GERD: Gastroesophageal Reflux Disease; HR: Heart Rate; MACE: Major Adverse Cardiovascular Events; NREM: Non-Rapid Eye Movement; $O_2$: Oxygen; REM: Rapid Eye Movement

Superscript numbers indicate the number of groups from which the value is significantly different. [+]Average sleep is self-reported and was available for only 1,822 samples.

Supplementary Table 8: Definitions of sleep variables

| Variable | Definition |
|---|---|
| **REM** | Rapid eye movement sleep stage |
| **NREM** | Non-rapid eye movement sleep stages |
| **N1** | Non-rapid eye movement sleep stage 1 |
| **N2** | Non-rapid eye movement sleep stage 2 |
| **N3** | Non-rapid eye movement sleep stage 3 |
| **AHI, total** | Apnea hypopnea index; Respiratory events per hour of sleep |
| **AHI, supine** | Respiratory events per hour of supine sleep |
| **AHI, off-supine** | Respiratory events per hour of off-supine sleep |
| **REM AHI** | Respiratory events per hour of REM sleep |
| **NREM AHI** | Respiratory events per hour of NREM sleep |
| **Obstructive apnea** | ≥10 seconds of ≥90% airflow reduction from pre-event baseline with continued respiratory effort |
| **Central apnea** | ≥10 seconds of ≥90% airflow reduction from pre-event baseline without respiratory effort |
| **Mixed apnea** | ≥10 seconds of ≥90% airflow reduction from pre-event baseline without respiratory effort for a portion of time and with respiratory effort for a portion of time |
| **Hypopnea** | ≥10 seconds of ≥30% reduction in the nasal transducer amplitude from pre-event baseline associated with either a 4% oxygen desaturation or a 3% oxygen desaturation or arousal* |
| **Central apnea index** | The number of central apneas per hour of sleep |
| **Arousal index** | Number of arousals per hour of sleep |
| **Epworth Sleepiness Scale** | Self-reported questionnaire that measures daytime sleepiness based on the likelihood of dozing off in various daily situations |
| **Total sleep time** | Total sleep time; expressed in minutes |
| **Sleep fragmentation** | The normalized power in the "fast" frequency range of the hypnogram's power spectral density (PSD), specifically transitions occurring faster than 10 minutes |
| **Sleep apnea-specific hypoxic burden** | The area under the curve of desaturations temporally related to respiratory events |
| **Delta Heart Rate** | Heart rate response; change in heart rate following respiratory events |
| **Lung 2 finger time** | Time between the end of a respiratory event and the lowest point of oxygen desaturation |

*Depending on insurer, as this is a clinical cohort

Supplementary Table 9: Logistic Regression (Odds Ratios, 95% Confidence Intervals) of Comorbidities among Patients from Different Risk Groups

| | Risk group | | | |
|---|---|---|---|---|
| **Disease** | RG**2 (n=1877)** | RG**3 (n=2867)** | RG**4 (n=1144)** | RG**5 (n=363)** |
| **Cardiovascular Risk Factors** | | | | |
| **Hypertension** | 1.02 (0.86-1.19) | 1.26 (1.10-1.45)(***) | 1.22 (1.00-1.50) | 1.36 (0.95-1.95) |
| **Diabetes Type 2** | 0.82 (0.70-0.96)(*) | 1.36 (1.20-1.53)(***) | 1.19 (1.01-1.40)(*) | 1.33 (1.03-1.73)(*) |
| **Hyperlipidemia** | 0.77 (0.65-0.90)(**) | 1.10 (0.96-1.25) | 1.00 (0.83-1.21) | 0.72 (0.53-0.98)(*) |
| **Obesity** | 0.81 (0.67-0.98)(*) | 2.49 (2.13-2.91)(***) | 1.68 (1.36-2.06)(***) | 1.59 (1.13-2.23)(**) |
| **Cardiovascular Disease** | | | | |
| **MACE** | 1.04 (0.86-1.26) | 1.53 (1.33-1.77)(***) | 1.64 (1.36-1.96)(***) | 1.41 (1.05-1.89)(*) |
| **Heart Failure** | 1.10 (0.83-1.46) | 1.81 (1.47-2.23)(***) | 1.86 (1.47-2.36)(***) | 2.11 (1.49-2.98)(***) |
| **Myocardial Infarction** | 1.02 (0.71-1.48) | 2.22 (1.71-2.89)(***) | 1.46 (1.07-1.99)(*) | 1.85 (1.19-2.87)(**) |
| **Ischemic Heart Disease** | 0.60 (0.28-1.25) | 2.21 (1.44-3.38)(***) | 1.64 (1.01-2.68)(*) | 2.00 (1.04-3.85)(*) |
| **Coronary Artery Disease** | 0.99 (0.79-1.25) | 1.38 (1.16-1.64)(***) | 1.38 (1.12-1.70)(**) | 1.16 (0.84-1.60) |
| **Atrial Fibrillation** | 0.89 (0.66-1.20) | 1.47 (1.19-1.82)(***) | 1.54 (1.21-1.96)(***) | 1.68 (1.18-2.41)(**) |
| **Stroke** | 1.03 (0.80-1.32) | 1.76 (1.47-2.12)(***) | 1.44 (1.16-1.81)(**) | 1.52 (1.08-2.13)(*) |
| **Ischemic Stroke** | 1.06 (0.79-1.41) | 1.74 (1.41-2.16)(***) | 1.45 (1.13-1.87)(**) | 1.53 (1.04-2.23)(*) |
| **Neurological Disorders** | | | | |
| **Mood Disorders** | 1.00 (0.87-1.16) | 1.59 (1.40-1.80)(***) | 1.58 (1.34-1.87)(***) | 1.67 (1.29-2.18)(***) |
| **Migraine** | 1.03 (0.86-1.24) | 1.34 (1.15-1.56)(***) | 0.96 (0.76-1.22) | 0.96 (0.64-1.42) |
| **Cognitive Impairment** | 1.15 (0.94-1.39) | 1.74 (1.50-2.03)(***) | 1.50 (1.24-1.83)(***) | 1.45 (1.06-1.97)(*) |
| **Epilepsy** | 1.09 (0.83-1.44) | 1.58 (1.27-1.98)(***) | 1.62 (1.21-2.18)(**) | 1.61 (1.00-2.58)(*) |
| **Other** | | | | |
| **GERD** | 0.98 (0.85-1.14) | 1.33 (1.18-1.49)(***) | 1.15 (0.98-1.35) | 1.02 (0.79-1.32) |
| **Acute Pain** | 0.58 (0.41-0.82)(**) | 2.71 (2.18-3.38)(***) | 1.57 (1.15-2.14)(**) | 1.25 (0.73-2.15) |
| **Chronic Pain** | 0.86 (0.74-1.01) | 2.45 (2.16-2.77)(***) | 1.61 (1.37-1.90)(***) | 1.45 (1.11-1.89)(**) |

Disease abbreviations: MACE: Major Adverse Cardiovascular Events; GERD: Gastroesophageal Reflux Disease

Significance levels: *p<0.05; **p<0.01; ***p<0.001

Supplementary Table 10: Logistic regression results for each disease at baseline

See attached file: logistic_summary_results.xlsx

Supplementary Table 11: Disease vocabulary used to parse the EMR data

See attached file: disease_vocabulary.xlsx