From Automated Scoring to Digital Biomarkers: Computational Methods Towards Precision Sleep Medicine

Inauguraldissertation
der Philosophisch-naturwissenschaftlichen Fakultät
der Universität Bern

vorgelegt von

Michal BECHNÝ

von Nový Jičín, Tschechien

Leiter der Arbeit:
Prof. Dr. Athina TZOVARA
Institut für Informatik

Dr. Francesca Dalia FARACI
Institute of Digital Technologies for Personalised Healthcare, SUPSI

From Automated Scoring to Digital Biomarkers: Computational Methods Towards Precision Sleep Medicine

Inauguraldissertation
der Philosophisch-naturwissenschaftlichen Fakultät
der Universität Bern

vorgelegt von

Michal BECHNÝ

von Nový Jičín, Tschechien

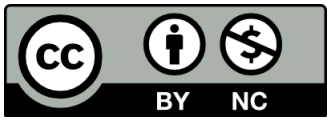
Leiter der Arbeit:
Prof. Dr. Athina TZOVARA
Institut für Informatik

Dr. Francesca Dalia FARACI
Institute of Digital Technologies for Personalised Healthcare, SUPSI

Von der Philosophisch-naturwissenschaftlichen Fakultät angenommen.

Bern, 24.10.2025

Der Dekan
Prof. Dr. Jean-Louis Reymond



Note on differing copyright licenses:

The above Creative Commons license applies to the thesis as a whole, except for Chapters 3, 6, and 7, which are subject to different copyright conditions as described below.

- **Chapter 3** — Based on the paper *"Framework for Algorithmic Bias Quantification and its Application to Automated Sleep Scoring"* by M. Bechny et al., published in the *Proceedings of the 11th IEEE Swiss Conference on Data Science (SDS 2024)*. © 2024 IEEE. This version reproduces the accepted manuscript with minor formatting and typographical corrections. It is included in the dissertation in accordance with the IEEE Author Rights Policy, which permits authors to use their accepted version in institutional repositories and theses, provided that the IEEE copyright notice and full citation are included. The official published version is available via IEEE Xplore (DOI: 10.1109/SDS60720.2024.00045). The Creative Commons license of this dissertation does not apply to this chapter.
- **Chapter 6** — Based on the contribution *"Unveiling Sleep Dysregulation in Chronic Fatigue Syndrome with and without Fibromyalgia Through Bayesian Networks"* by M. Bechny et al., published in the *Proceedings of the 23rd International Conference on Artificial Intelligence in Medicine (AIME 2025)*, Lecture Notes in Artificial Intelligence, Springer Nature Switzerland AG. © 2025 Springer Nature Switzerland AG. This version reproduces the accepted manuscript with minor formatting and typographical corrections. It is included in the dissertation in accordance with the Springer Nature Licence-to-Publish agreement, which permits authors to reuse their contribution in a doctoral thesis and to make that thesis publicly available in an institutional repository, provided that proper acknowledgement and citation are given. The official published version will be available via Springer Nature at DOI: 10.1007/978-3-031-95838-0_4. The Creative Commons license of this dissertation does not apply to this chapter.
- **Chapter 7** — Based on the preprint *"Sleep-Stage Dynamics Predict Current Sleep-Disordered Breathing and Future Cardiovascular Risk"* by M. Bechny et al., posted on *medRxiv* on August 1, 2025 (DOI: 10.1101/2025.07.31.25332545). © 2025 The Author(s). The copyright holder for this preprint is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a Creative Commons Attribution (CC BY 4.0) International License. This chapter reproduces and extends the content of the preprint. The Creative Commons license of this dissertation applies equally to this chapter.

Abstract

From Automated Scoring to Digital Biomarkers: Computational Methods Towards Precision Sleep Medicine

Michal BECHNÝ, Ph.D. in Computer Science

Universität Bern, 2025

Sleep, together with diet and physical activity, is one of the fundamental pillars of health. Sleep disorders are highly prevalent, with rising incidence particularly among younger and economically disadvantaged populations, and they are closely linked to neurological, psychiatric, metabolic, and cardiovascular conditions. The clinical gold standard for assessing sleep is polysomnography (PSG), which records multiple biosignals such as brain activity, breathing, and movement during the night. Following established guidelines, every 30-second segment is scored by trained experts into one of five sleep stages, producing a stage-by-stage sequence called a hypnogram. From this representation, standard sleep metrics are calculated and combined with indices of breathing- or movement-related events to support diagnostics. While central to diagnosis, PSG is costly and labour-intensive, with manual scoring alone requiring up to two hours of expert time. With the growing use of Artificial Intelligence (AI), automated sleep scoring (ASS) powered by modern machine- and deep-learning algorithms achieves human-level agreement of 75–85%, but remains limited by the inter-rater variability inherent in training labels. This limits performance and requires mechanisms for effective human–AI collaboration. The first branch of this thesis (Chapters 2–4) addresses these challenges by developing methods for clinical integration of ASS, including uncertainty-guided oversight and a flexible statistical framework to quantify algorithmic bias. These approaches aim to support efficient physician review while promoting fairness, transparency, and reliability in clinical deployment. After improving the ASS process, the second branch (Chapters 5–7) focuses on deriving novel digital biomarkers from *sleep-stage dynamics*—an underutilised aspect of PSG with potential to capture subtle physiological signatures. Using clinical, observational, and prospective datasets, we investigated their value in sleep-disordered breathing, chronic fatigue and pain syndromes, and long-term cardiovascular risk. To address confounding and build predictive models, we employed causal inference methods, Bayesian networks, and forest-based classification and survival models, systematically examining the effects and risk profiles associated with sleep-stage transitions. These studies revealed disorder-specific alterations and showed that sleep-stage dynamics can also predict future cardiovascular events. Together, the findings demonstrate that sleep-stage dynamics represent a promising class of digital biomarkers that extend standard PSG metrics, improve risk assessment, and—when combined with wearable technology in the future—may enable unobtrusive yet sensitive long-term monitoring of diverse conditions.

Acknowledgements

I am very grateful to my PhD advisors, Athina Tzovara from the Cognitive Computational Neuroscience (CCN) group at the University of Bern and Francesca Faraci from the Biomedical Signal Processing group at the Institute of Digital Technologies for Personalised Healthcare (MeDiTech) at SUPSI in Lugano. Thank you for the time you dedicated to me, the knowledge you shared, and the opportunities you made possible. I especially thank Francesca for securing the funding that allowed me to pursue this PhD, present my work at international conferences, and for her personal support during challenging moments.

My work with data from Inselspital, University Hospital of Bern, would not have been possible without Julia van der Meer, whose support in understanding the data, along with her constructive feedback even under tight deadlines, proved invaluable. I also thank Markus Schmidt and Claudio Bassetti for offering clinical perspectives on my work.

I am grateful to Akifumi Kishi, who supervised me during my research stay in Japan and gave me valuable input on sleep dynamics, as well as being a wonderful guide in the Land of the Rising Sun. I also thank Yasuhiro Tomita for his helpful feedback on the link between sleep and cardiovascular disease, and for the memorable dinner we shared in Tokyo.

Many thanks to Marco Scutari, for his openness in discussing methods, the knowledge he passed on, the academic connections he introduced me to, and his generous support.

I owe thanks as well to all my colleagues in Lugano, Bern, and Tokyo for their feedback on my work and the good moments we shared over lunch and outside work. I also thank my former university teachers, especially the team at the Institute of Applied Statistics at JKU Linz, and my former research colleagues at the Software Competence Center Hagenberg in Austria, who passed on their knowledge and helped me at the start of my career.

I gratefully acknowledge the University of Bern for supporting my research stay in Japan and my attendance at the World Sleep Congress in Singapore, both of which were very valuable experiences as a young researcher. I also thank my hometown, Nový Jičín, for awarding me a scholarship for international students that helped cover some of my conference expenses, and the Swiss Society for Sleep Research, Sleep Medicine and Chronobiology (SSSSC) for a travel grant to the European Sleep Congress in Seville.

Finally, my warmest thanks go to my beloved partner Radka, my family, and my friends. I thank Radka for her support, patience, and respect, which have been the driving force of our long-standing relationship. I am especially grateful to my grandparents, whose example has been one of my greatest teachers. I also thank my siblings for always being there for me, and above all, my mother, for her endless love and understanding.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Sleep, alongside physical activity and diet, is recognised as one of the three fundamental pillars of human health. The connection between sleep and well-being has been acknowledged for millennia. One of the earliest preserved references comes from ancient Greece, attributed to Hippocrates, often regarded as the father of medicine:

> *"Both sleep and insomnolency, when immoderate, are bad."*
> —Hippocrates (c. 400 BCE)

Despite its age, this quote remains remarkably relevant today. Hippocrates did not reduce the complexity of sleep needs to a simple *"more is better"* argument. Instead, he implied, likely deliberately without specifying a quantity, that there exists an optimal amount of sleep, and that both insufficient and excessive sleep may be harmful.

In this short aphorism, Hippocrates anticipated many modern efforts to quantify the ideal duration and structure of sleep. Yet much of contemporary thinking, both in popular discourse and in scientific literature, still assumes that *"more"*—whether more total sleep, more deep sleep, or more rapid eye movement (REM) sleep—is inherently beneficial. Far fewer studies consider the possibility of a U-shaped relationship, in which both extremes are detrimental and the optimum lies somewhere in between. This concept, intuitive even two and a half millennia ago, remains underutilised in modern research, perhaps because identifying such an optimum, which likely varies between individuals and population subgroups, is computationally and practically challenging.

The endurance of Hippocrates's words suggests that sleep disturbances were already a recognised health concern in ancient times. Whether their prevalence in ancient Greece was comparable to today is unknown, but modern epidemiological data are sobering: in Switzerland, for example, about one-third of the population reports sleep disorders: 7% pathological and 26% moderate, with nearly half (48%) waking multiple times during the night, either frequently or occasionally [1]. The causes and distribution of sleep problems have shifted considerably over time: disorders related to physical inactivity and obesity, such as obstructive sleep apnea, or those linked to excessive exposure to artificial light before bedtime, were likely rare in ancient societies but are increasingly common today.

While our intuitive understanding of sleep dates back millennia, substantial scientific progress has been achieved only in recent decades. Advances in neuroscience, physiology, and biomedical engineering have deepened our understanding of sleep architecture and disorders, while the emergence of sleep medicine as a clinical discipline has been enabled by technological innovation. Specialised sleep laboratories, formal diagnostic criteria, and advanced monitoring tools—ranging from gold-standard polysomnography to modern wearable devices—now allow for precise diagnosis, long-term monitoring, and detailed assessment of sleep-related behaviours.

This dissertation aims to contribute to the field of sleep medicine not by directly improving the reader's sleep (although certain sections may have that side effect), but by applying quantitative methods to advance the computational assessment and interpretation of sleep. It bridges the domains of *computer science* and *clinical sleep medicine*, demonstrating how algorithmic approaches, data-driven modelling, and machine learning can be applied to address practical problems in healthcare effectively.

The contributions of this thesis fall into two main thematic branches:

1. **Clinical integration of algorithms for automated sleep scoring (ASS)**, with the potential to streamline clinical workflows, reduce the cost of sleep studies, and promote algorithmic fairness in clinical decision-making.
2. **Quantification of novel digital biomarkers derived from sleep-stage dynamics**, to enhance physiological insight into sleep disorders, support their diagnosis, and investigate their associations with long-term health outcomes, particularly in the cardiovascular domain.

In the sections that follow, we introduce the clinical sleep study, polysomnography (PSG), and its outputs, including the process of sleep scoring. We then examine the challenges of (automated) sleep scoring, such as inter-scorer disagreement, and the potential of scoring outputs for biomarker derivation and diagnostics. This contextual foundation is followed by an overview of the six manuscripts that form the core of this dissertation and their interconnections. The subsequent chapters present each manuscript in full, while the final chapter synthesises the findings, explores their implications for sleep medicine, and discusses their limitations.

1.1 Clinical Sleep Study (Polysomnography)

Polysomnography (PSG) is the gold standard for the objective assessment of sleep physiology and the diagnosis of sleep disorders. These disorders, such as insomnia, sleep-disordered breathing (SDB, including obstructive and central sleep apnea), hypersomnia, parasomnias, movement-related sleep disorders, and circadian rhythm disturbances, affect a substantial portion of the population and are closely linked to cardiovascular, metabolic, and mental health outcomes [2]–[7].

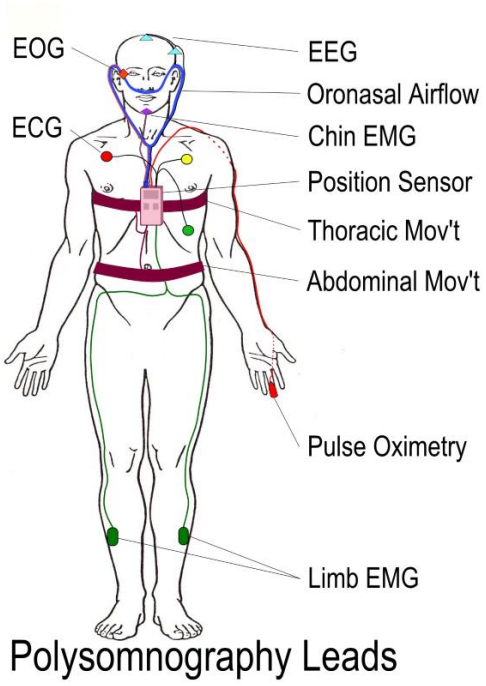
A typical overnight PSG combines neurophysiological, cardiorespiratory, and musculoskeletal recordings to provide detailed insight into sleep architecture and physiology. It enables both sleep staging and the detection of physiological or pathological events such as apneas, periodic limb movements, and cortical arousals [8], [9].

1.1.1 Polysomnographic Acquisition

In a standard clinical PSG setup (Figure 1.1), the following physiological signals are recorded overnight in a controlled laboratory setting [8], [10], [11]:

- **Electroencephalography (EEG)** – to monitor brain activity and determine transitions between wakefulness and individual sleep stages, using electrodes placed according to the international 10–20 system;
- **Electrooculography (EOG)** – to capture horizontal and vertical eye movements, essential for identifying REM sleep;
- **Electromyography (EMG)** – to monitor muscle tone and activity, particularly in the chin (submental region) and lower limbs (tibialis anterior), relevant for detecting REM atonia and movement-related sleep disorders;
- **Respiratory effort and airflow** – measured using thoracic and abdominal belts (typically respiratory inductance plethysmography) and nasal pressure transducers or thermistors;
- **Pulse oximetry (SpO_2)** – to detect oxygen desaturation events associated with sleep-disordered breathing;
- **Body position sensors** – to track posture and positional dependency of respiratory events;
- **Optional channels** – including electrocardiography (ECG) for heart rate and rhythm monitoring; snore microphones; and additional limb EMG channels for periodic limb movement detection.

Figure 1.1: Typical polysomnographic setup.



Notes: Taken from [12].

Signals are digitised and stored at high resolution (200–512 Hz) for manual or automated analysis. In-lab PSGs often include synchronised infrared video and audio to assist with behavioural assessment and artifact detection. Although requiring extensive patient instrumentation and technician oversight, this multichannel setup offers a comprehensive view of sleep architecture and related pathologies.

1.1.2 Manual Sleep Scoring

Sleep stages are traditionally assigned in 30-second epochs by visual inspection of EEG, EOG, and EMG signals, following the American Academy of Sleep Medicine (AASM) criteria [8]. The five standard vigilance states with unique physiology are [8], [9], [13]:

- **Wake (W)** – Characterized by alpha activity (8–12 Hz) over the occipital region during eyes-closed rest, transitioning to beta activity (13-30 Hz) with eyes open or increased alertness. Muscle tone is high, and frequent eye movements are present. In healthy individuals, wakefulness typically accounts for 5–10% of the total PSG recording.
- **Stage N1** – The lightest sleep stage, marked by low-amplitude mixed-frequency EEG (theta 4–7 Hz), slow rolling eye movements, and reduced chin EMG tone. N1 is a brief transitional stage comprising ∼5% of sleep in healthy adults.
- **Stage N2** – Defined by the appearance of sleep spindles (= about 1 second short, 11–16 Hz oscillations) and K-complexes (sharp negative wave followed by a slower positive component). Eye movements cease, and muscle tone decreases. N2 typically follows N1 and represents 45–55% of total sleep time.
- **Stage N3** – Also called slow-wave sleep (SWS), exhibits high-amplitude low-frequency delta waves (0.5–3 Hz). Muscle tone is low, and arousal thresholds are highest. N3 comprises 15–25% of sleep, mainly in the early part of the night.
- **REM sleep** – Identified by low-amplitude mixed-frequency EEG, rapid eye movements, and near-complete muscle atonia. REM typically follows N2 and accounts for 20–25% of sleep, increasing in duration toward the later part of the night.

In healthy adults, these stages cycle every 90–120 minutes, following a pattern of: (W → N1 → N2 → N3 → N2 → REM → N1/W), with N3 declining and REM increasing across the night [9], [13]. Awakenings are more likely from N1, N2, or REM [9], [13].

Sleep stages and their composition (i.e., *macrostructure*) provide clinicians with valuable information about possible sleep disturbances or pathologies. Manual scoring, however, is both time-consuming and prone to variability. Inter-scorer agreement typically ranges between 75–85%, depending on dataset complexity (e.g., healthy vs. clinical populations; pediatric vs. adult recordings) and the level of scorer training [14], [15]. Scoring a single night's PSG generally takes 1–2 hours of focused work by a trained technologist [16].

1.1.3 PSG-Derived Sleep Metrics

The PSG recording begins with the *lights-off* time, marking the intended onset of the study and start of biosignals measurements, and ends with the *lights-on* time. The difference between lights-off and lights-on defines the *total recording duration* or *time in bed*. During this interval, sleep stages are scored epoch by epoch, yielding a temporal sequence known as a *hypnogram*, providing a visual summary of sleep architecture across the night. The hypnogram allows for the extraction of sleep macrostructure metrics, which are, alongside with interpretation of raw biosignals and derived indices, used as standard clinical markers in the evaluation of sleep quality and the diagnosis of sleep disorders [8].

Hypnogram-Derived Metrics include:

- **Total Sleep Time (TST)** – total duration scored in sleep stages (N1, N2, N3, REM); excludes epochs scored as W.
- **Sleep Efficiency (SE)** – ratio of TST to total time in bed (from lights-off to lights-on), expressed as a percentage.
- **Sleep Latency (SL)** – time from lights-off to the first epoch scored as sleep (usually N1).
- **REM Latency** – time from sleep onset (first non-W epoch) to first REM epoch.
- **Wake After Sleep Onset (WASO)** – total time spent awake after initial sleep onset and before final awakening.
- **Sleep Stage Percentages** – relative share of TST spent in N1, N2, N3, and REM sleep.
- **Number of Awakenings per Hour** – rate of transitions from sleep (N1, N2, N3, REM) to wakefulness (W), used to quantify sleep continuity or fragmentation.
- **Number of Stage Transitions per Hour** – rate of stage switching, reflecting overall sleep stability or fragmentation.

Biosignal-Derived Indices, derived from EEG, EOG, EMG, and respiratory channels, include particularly [8]:

- **Arousal Index** – Number of EEG-defined cortical arousals ($\geq$3 s) per hour of sleep, following $\geq$10 s of stable sleep.
- **REM Density** – The number of eye movements per minute of REM, often associated with mood disorders such as depression.
- **Periodic Limb Movement Index (PLMI)** – Number of periodic limb movements (0.5–10 s each, spaced 5–90 s apart) per hour, detected via EMG from the anterior tibialis muscle.
- **Apnea-Hypopnea Index (AHI)** – Number of apneas ($\geq$10 s airflow cessation) and hypopneas ($\geq$30% airflow reduction + $\geq$3% desaturation or cortical arousal) per hour of sleep.
- **Oxygen Desaturation Index (ODI)** – Number of $\geq$3% oxygen desaturations ($\geq$10 s) per hour from baseline, measured via pulse oximetry.

Impact of disorder, age, and gender on sleep

Hypnogram-derived metrics and biosignal-based indices are influenced by age and gender, even in generally healthy individuals, and are further altered in the presence of sleep and other disorders or medications. With ageing, TST decreases, fragmentation increases (WASO, awakenings, stage transitions), and N3 proportion declines markedly [9], [17], [18]; REM sleep may also become less stable [17], [19]. Gender differences are also evident: females tend to have longer TST, higher SE, more N3, and shorter SL, despite reporting more sleep complaints [18], [20], [21]; males show greater fragmentation and steeper age-related declines in deep sleep [18], [21]. These variations must be taken into account to distinguish typical ageing and sex effects from pathological alterations in clinical interpretation.

Sleep disorders and other comorbidities alter sleep architecture and associated metrics in distinct ways. Insomnia is characterized by reduced TST, SE, and prolonged SL/WASO, reflecting difficulties with sleep initiation and maintenance [22]. Hypersomnia disorders such as narcolepsy feature shortened SL and REML, along with frequent sleep-onset REM periods (SOREMPs) [23], [24]. Sleep-disordered breathing, particularly obstructive sleep apnea (OSA), is typically marked by reduced N3 and REM sleep, elevated WASO, and frequent arousals and awakenings, accompanied by increased AHI and ODI [25]–[27]. REM sleep behavior disorder (RBD) involves loss of REM atonia, abnormal stage transitions, and altered REM structure [28]–[30]. Depression is often associated with shortened REM latency, increased REM percentage, and fragmented sleep architecture [31], [32]. Neurodegenerative, metabolic, and renal disorders commonly present with reduced SE and TST, increased WASO, and loss of N3 and REM sleep [33]–[36]. Finally, reduced N3 and REM sleep, diminished delta activity, and abnormal total sleep duration have been consistently associated with increased cardiovascular morbidity and all-cause mortality [37]–[44].

Altogether, PSG-derived sleep metrics not only enable the diagnosis of sleep disorders but also provide a valuable reflection of broader physiological and pathological processes across the lifespan.

1.1.4 Cost and Accessibility of PSG Across Healthcare Systems

Polysomnography (PSG) is a resource-intensive diagnostic procedure, and its cost, availability, and reimbursement vary considerably across healthcare systems. For example, in the United States (U.S.), the total cost of an attended, in-laboratory PSG typically ranges from $1,000 to more than $7,000, depending on geographic region, clinical setting, and whether associated services—such as pre-study consultation and post-study interpretation—are included [45], [46]. The average market price is estimated at approximately $3,300 per study [45]. In Switzerland, PSG is available at most public and private sleep centers at a cost ranging from 1,000 to 3,500 CHF, generally reimbursed under compulsory health insurance and subject to applicable deductibles and co-payments [47]. Across Europe, access to accredited sleep laboratories is generally well-established, although regional differences in availability and waiting times persist. Home-based sleep studies offer a lower-cost alternative but with reduced diagnostic resolution. The main cost drivers of PSG include technician labor for overnight monitoring and manual scoring, physician time for interpretation, and the amortization of equipment and laboratory infrastructure. A standard PSG involves 1–2 hours for patient preparation, 6–8 hours of overnight recording, and 1–2 hours for post-acquisition scoring and analysis [16].

Cost and limited availability remain significant barriers to broader clinical implementation of PSG, particularly for early screening and routine assessment.

1.2 Automated Sleep Scoring: Potential and Challenges

1.2.1 Background and Historical Development

Manual sleep staging of PSG recordings is time-consuming, costly, and requires trained scorers. To reduce this burden, automated sleep scoring (ASS) has been explored since the

1960s [48]. The aim is to algorithmically replicate the AASM rules [8], assigning each 30-second PSG epoch to one of five vigilance states—W, N1, N2, N3, or REM—thus framing the task as multiclass classification.

Early approaches were rule-based or statistical, aiming to mimic formal scoring criteria into handcrafted logic and operating on small datasets [49]–[51]. With the development of the broader field of Artificial Intelligence (AI), these were followed by classical machine-learning methods that relied on manually extracted features mapped to target stages in a more data-driven manner [52]–[58]. In the past decade, deep-learning (DL) architectures—convolutional, recurrent, and more recently transformer-based—have dominated the field [48], [59]–[64]. These models learn stage-discriminative representations directly from raw biosignals, eliminating the need for expert-defined features.

Despite these technical advances, generalization of ASS algorithms on unseen data typically saturates at 75–85% for common performance metrics such as accuracy, macro-F1, or Cohen's κ—aligned with the average human inter-scorer agreement [14], [15], [48], [65]–[69].

To better encode labelling uncertainty, a part of ASS research exploits multi-scorer datasets, in which each PSG is independently annotated by several experts [70]–[73]. These studies show that modern ASS models can match, or even outperform, individual scorers when compared to a consensus reference, demonstrating the potential of ASS systems to provide robust and clinically useful outputs. However, collecting such datasets is resource-intensive, limiting their widespread availability.

Cross-study comparisons of different ASS systems are further complicated by differences in datasets, preprocessing, training pipelines, and evaluation metrics. Recent benchmarking initiatives have addressed this by harmonizing datasets and training protocols, aiming to improve reproducibility and representativeness of study comparisons [73], [74]. These efforts suggest that, when trained on the same data, sufficiently capable architectures (e.g., U-Net, DeepResNet, transformer) tend to converge toward similar levels in performance metrics [73] and trends, when associated with demographics and clinical variables.

As the field advances, key priorities include the creation of large, representative, open-access (and ideally multi-scorer) databases; standardized benchmarking protocols; and a deeper understanding of model generalization. These steps are essential for safe and effective deployment of ASS in clinical practice.

1.2.2 Current Clinical Use and Regulatory Considerations

Despite major technological advances and the proliferation of ASS solutions, manual scoring by trained physicians or technicians remains the clinical standard. This persistence reflects a combination of technical, ethical, and regulatory constraints that currently prevent full automation of sleep-scoring within clinical practice.

From a technical perspective, generalization remains the central challenge. Supervised ASS models are trained on human-labelled PSG data, which are inherently noisy due to scorer subjectivity, signal artefacts, and inter- as well as intra-subject variability. Inter-scorer agreement for sleep staging typically ranges from 75–85%, depending on dataset composition and clinical context [14], [15], [48], [65]–[69]. This variability stems from multiple sources: subjective interpretation of biosignals; complexity in applying the AASM rules; demographic and clinical heterogeneity (e.g., due to age, sex, comorbidities); and possible ambiguities in the rules themselves. Historical changes in scoring criteria, such as the transition from the R&K system to the AASM guidelines [75]–[77], further compound inconsistency in historical datasets.

Moreover, algorithm performance often varies systematically across subpopulations, reflecting differences in sleep architecture and biosignal characteristics driven by factors such as age and health status [18], [58], [78], [79].

As a result, the noise in the sleep scoring labels, primarily due to inter-rater variability, places a theoretical performance ceiling on ASS models trained on large heterogeneous data, typically between 75–85% in performance metrics [80], [81]. Moreover, the performance of ASS algorithms often varies systematically across subpopulations, reflecting differences in sleep architecture and biosignal characteristics driven by factors such as age and health status [18], [58], [78], [79].

A further limitation is that most ASS research focuses on stage-level performance metrics while overlooking clinical validity and diagnostic utility. Only a small subset of studies, primarily in the context of at-home PSG systems seeking regulatory approval or certification, evaluate whether model-derived scoring supports accurate computation of downstream, hypnogram-derived, clinical markers (e.g., WASO, TST), and quantify associated error rates [82]–[84]. Without such evidence, the real-world clinical utility of ASS remains difficult to assess.

Ethical and Legal Considerations:

As software-based medical technologies, often incorporating AI, ASS systems fall under evolving legal and regulatory frameworks governing AI in healthcare. Relevant instruments include the EU AI Act (Regulation (EU) 2024/1689), the EU Medical Device Regulation (MDR, Regulation (EU) 2017/745), the Swiss Medical Devices Ordinance (MedDO, SR 812.213), and the U.S. FDA's guidance on AI/ML-enabled medical devices. These frameworks embed ethical principles such as transparency, fairness, and accountability, while mandating human oversight for AI systems involved in high-stakes clinical decision-making [85]–[87].

In ASS, AI-based algorithms are constrained by inherent inter-scorer variability, leading to typical disagreement rates of 15–25% between model outputs and human annotations [14], [15], [65], [66], [88]. Given this variability, it is often impossible to determine whether a specific discrepancy reflects an algorithmic error or a difference in human interpretation. Yet, in the clinical setting, legal and medical accountability remains entirely with the physician. Consequently, human-labelled annotations persist as the definitive reference standard, both in regulation and practice. In line with this, current regulatory frameworks mandate a cautious approach to the adoption of AI-based software tools, including ASS: such systems must complement rather than replace expert judgment, and must meet rigorous requirements for explainability, fairness, and post-deployment surveillance [89]–[91].

Beyond performance, regulators and researchers increasingly stress the importance of addressing algorithmic fairness and equity in healthcare AI, to prevent the amplification of existing health disparities [89], [90], [92]. This consideration is particularly relevant for ASS systems, where reliance on observational data with uneven distributions of demographic and clinical variables can lead to unequal model performance across patient subgroups, thereby directly affecting their clinical outcomes.

While much research focuses on maximizing ASS performance—which, given label noise, is already at its theoretical ceiling [15], [65], [88]—few studies address regulatory compliance or effective clinical integration. In particular, limited work evaluates how ASS tools can be embedded into efficient workflows that enable meaningful interaction with human experts [93], [94], or validates the diagnostic utility of prediction-derived clinical markers [95]–[97].

Going forward, research efforts must balance technical optimisation with practical deployment. Priorities include the creation of representative training datasets, the standardisation of benchmarking practices [71], [73], and the design of clinically integrated ASS tools that meet regulatory, ethical, and usability requirements.

1.3 Structure of the thesis

Motivated by the importance of sleep as one of the pillars of health, the cost and limited availability of polysomnography (PSG), and the potential of the large volume of data it produces, this dissertation presents six closely related first-author studies by Michal Bechny, PhD candidate in Computer Science at the University of Bern. Together, these works range from uncertainty quantification and bias analysis in predictive ASS algorithms to causal inference and explainable machine learning for biomarker discovery, with a shared focus on improving the computational assessment of sleep and its integration into clinical workflows. For clarity, the research is organised into two main conceptual branches.

The first part of the thesis focuses on the design and application of computational methods for the effective **Integration of ASS into Clinical Practice**. It primarily addresses the

general ethical and legal mandates for introducing AI-based (software) solutions into healthcare, using ASS as a case study. Specifically, the presented studies aim to develop an efficient *human-in-the-loop* pipeline, to ensure human oversight, and to promote AI fairness by introducing a general framework for *detecting and quantifying algorithmic bias*. This thematic line is covered in:

- **Chapter 2** — *Bridging AI and Clinical Practice: Integrating Automated Sleep Scoring Algorithm with Uncertainty-Guided Physician Review* [95], focusing on establishing human oversight in ASS by quantifying prediction uncertainty, published in *Nature and Science of Sleep*. The work compares several approaches, including a proposed auxiliary long short-term memory (LSTM) confidence neural network, specifically designed for time-series PSG data, estimating predictive uncertainty in DL-based algorithm U-Sleep, and their use in establishing and streamlining human oversight.

- **Chapter 3** — *Framework for Algorithmic Bias Quantification and its Application to Automated Sleep Scoring* [96], presenting a method to quantify systematic deviations (i.e., *biases*) in predictive algorithms, published as a short paper at the *2024 11th IEEE Swiss Conference on Data Science (SDS)*. The study introduces a Generalized Additive Models for Location, Scale, and Shape (GAMLSS)-based framework to characterise the distribution of errors in algorithmic predictions, conditioned on sensitive attributes of interest, such as demographic and clinical characteristics. The approach is illustrated on the U-Sleep algorithm [59], [60] in the context of wakefulness detection.

- **Chapter 4** — the follow-up manuscript *Beyond Accuracy: A Framework for Evaluating Algorithmic Bias and Performance, Applied to Automated Sleep Scoring* [97], published in *Scientific Reports*, extends the GAMLSS framework by quantifying the conditional distribution of performance metrics and presents it on two state-of-the-art ASS algorithms. Bias is evaluated in a wide range of clinical hypnogram-derived markers and several performance metrics. The clinical utility of possibly biased ASS predictions for diagnostics is also discussed.

After enhancing the ASS process, the second part of the thesis focuses on the design and quantification of novel **Digital Biomarkers from Sleep-Stage Dynamics** to characterise current health status, understand the impact of different disorders, and quantify the risk of long-term health outcomes, using explainable machine learning methods. The characteristics of sleep-stage dynamics are, despite evidence of being capable of capturing detailed physiological signatures (cf. [98]–[114]), still underutilised in standard clinical PSG studies. Current PSG reports typically include only coarse parameters such as total or hourly awakenings, or the overall rate of stage transitions. Motivated by this underuse and the potential of sleep-stage dynamics for diagnostics and risk assessment, this branch of the thesis presents the following studies:

- **Chapter 5** — *Novel Digital Markers of Sleep Dynamics: Causal Inference Approach Revealing Age and Gender Phenotypes in Obstructive Sleep Apnea* [115], published in *Scientific Reports*, introduces a causal meta-learner approach for personalised digital markers derived from sleep-stage dynamics and applies it to the (comorbid) apnea use-case. The work links the matrix of sleep-stage transition proportions to established hypnogram-based clinical parameters, and shows how to exploit it to derive new markers using propensity and outcome models with logistic and Dirichlet regression, adjusted for clinical confounders.

- **Chapter 6** — *Unveiling Sleep Dysregulation in Chronic Fatigue Syndrome with and without Fibromyalgia Through Bayesian Networks* [116], published as a full-paper at the *23rd International Conference on Artificial Intelligence in Medicine (AIME 2025)*, quantifies sleep-stage dynamics using Bayesian networks (BN). Using a strictly controlled dataset, the study quantifies the impact of chronic fatigue syndrome (CFS) and its interaction with fibromyalgia (FM) on sleep dynamics. The work exploits BN for both diagnostic classification of health states and estimation of causal effects, providing evidence for clinical differentiation and suggesting novel diagnostic markers.

- **Chapter 7** — *Sleep-Stage Dynamics Predict Current Sleep-Disordered Breathing and Future Cardiovascular Risk* [117], under review in *Scientific Reports*, applies explainable forest-based machine learning models to demonstrate that sleep-stage dynamics can not only diagnose current condition (SDB) but also predict the risk of future cardiovascular events, reflecting the central role of sleep in human health. Interpretability techniques based on partial dependence effects reveal novel markers and risk profiles for both SDB and long-term cardiovascular outcomes.

Collectively, the presented works draw on statistical modelling, causal inference, and machine learning to advance both the methodological and clinical frontiers of sleep research, with a consistent focus on reliability, fairness, and practical utility. The thesis concludes with **Chapter 8**, which summarises the main findings and contributions of the presented studies, as well as their limitations.

Chapter 2

Bridging AI and Clinical Practice: Integrating Automated Sleep Scoring Algorithm with Uncertainty-Guided Physician Review

Abstract

Purpose: This study aims to enhance the clinical use of automated sleep-scoring algorithms by incorporating an uncertainty estimation approach to efficiently assist clinicians in the manual review of predicted hypnograms, a necessity due to the notable inter-scorer variability inherent in polysomnography (PSG) databases. Our efforts target the extent of review required to achieve predefined agreement levels, examining both in-domain (ID) and out-of-domain (OOD) data, and considering subjects' diagnoses. **Patients and Methods:** A total of 19,578 PSGs from 13 open-access databases were used to train U-Sleep, a state-of-the-art sleep-scoring algorithm. We leveraged a comprehensive clinical database of an additional 8832 PSGs, covering a full spectrum of ages (0– 91 years) and sleep-disorders, to refine the U-Sleep, and to evaluate different uncertainty-quantification approaches, including our novel confidence network. The ID data consisted of PSGs scored by over 50 physicians, and the two OOD sets comprised recordings each scored by a unique senior physician. **Results:** U-Sleep demonstrated robust performance, with Cohen's kappa (κ) at 76.2% on ID and 73.8–78.8% on OOD data. The confidence network excelled at identifying uncertain predictions, achieving AUROC scores of 85.7% on ID and 82.5-85.6% on OOD data. Independently of sleep-disorder status, statistical evaluations revealed significant differences in confidence scores between aligning vs discording predictions, and significant correlations of confidence scores with classification performance metrics. To achieve $\kappa \geq 90\%$ with physician intervention, examining less than 29.0% of uncertain epochs was required, substantially reducing physicians' workload, and facilitating near-perfect agreement. **Conclusion:** Inter-scorer variability limits the accuracy of the scoring algorithms to $\sim$80%. By integrating an uncertainty estimation with U-Sleep, we enhance the review of predicted hypnograms, to align with the scoring taste of a responsible physician. Validated across ID and OOD data and various sleep-disorders, our approach offers a strategy to boost automated scoring tools' usability in clinical settings.

Keywords:
Automated Sleep Scoring, Uncertainty Quantification, Explainable AI, Polysomnography, Sleep Medicine

2.1 Introduction

Sleep, often dubbed as the third pillar of health alongside diet and exercise, plays a critical role in our well-being. Polysomnography (PSG), a comprehensive sleep monitoring technique, captures detailed biosignals – primarily the electroencephalogram (EEG), the electrooculogram (EOG), and the electromyogram (EMG). Adhering to the guidelines of the American Academy of Sleep Medicine (AASM) [8], physicians score PSG recordings into specific sleep stages, on 30-second windows (epochs). Such structured scoring, called a *hypnogram*, divides sleep into five distinct stages: W, REM, N1, N2, and N3, each representing a unique physiological state [13]. The proportions of sleep stages, as well as patterns in their transitions, are basic indicators of sleep health [118], [119], and also biomarkers of certain disorders [27], [104], [120].

While manual scoring remains the gold standard, the procedure may be labor-intensive, often demanding up to 2 hours for a comprehensive evaluation of a single PSG recording [16]. Research into automatic sleep scoring, which aims to support the manual scoring of physicians by computational algorithms, dates back to the 1960s [48]. Recent advancements in Artificial Intelligence (AI) have significantly improved automatic scoring solutions, especially those based on Machine and Deep Learning (ML/DL) methodologies. Notably, the U-Sleep algorithm introduced by Perslev et al. [59], and further investigated by Fiorillo & Monachino et al. [60], stands at the forefront due to its balance between performance rivaling human scorers and the diversity of its training data.

Supervised automated sleep scoring algorithms can reach considerable performance but are to-date not able to overcome an intrinsic problem. The different interpretations of AASM scoring standards by physicians result in an inter-scorer agreement of about 76% [15], [65], [88]. This human-based variability in the annotations introduces approximately 20% noise-level, technically limiting the performance of scoring algorithms optimized in a supervised way, as the ability of an AI algorithm can hardly be better than the quality of its training data. Consequently, despite the breadth of training databases available, the ceiling for ML/DL model generalizability is limited by this prevailing inter-scorer agreement. Therefore, despite the technological advancements AI has brought to sleep scoring, physicians – who are still irreplaceable and responsible for clinical decisions – must subject the predicted hypnograms to a thorough review and compare whether the algorithm-proposed predictions are consistent with their personal interpretation of patterns present in the original PSG biosignals. While some level of error in sleep-scoring models is deemed clinically acceptable [121], the review process of predicted hypnograms can be time-consuming and costly. Specifically, if physicians lack prior insights into problematic segments of the biosignal, the review might be as resource-intensive as conducting manual scoring without any algorithmic assistance.

Given the limits posed by inter-scorer variability, a subset of research has pivoted towards quantifying prediction uncertainty to elevate model performance by enabling review of the least confident predictions. Such semi-automated approaches combining predictions proposed by algorithms with physician's expertise represent a promising solution for integration of sleep scoring tools in clinical settings [48]. Van Gorp et al. delved into the theoretical aspects of such (un)certainty [93]. Kang et al. advanced this notion by proposing an uncertainty detection mechanism via Shannon's entropy of the softmax output of a statistical classifier [122]. By allowing physicians to correct uncertain predictions, they managed to substantially enhance the agreement (κ-score) between classifier and physician's scoring taste. In the realm of DL-based algorithms, Fiorillo et al. employed a query procedure targeting a predetermined percentage of the most uncertain predictions based on the maximum and variance of the softmax output [123]. Hong et al. presented a novel method, Dropout-Correct-Rate, and showcased its potential to boost model performance with targeted human review [94]. Meanwhile, Phan et al. utilized a transformer-based sleep scoring model and identified uncertain epochs through normalized entropy scores, demonstrating that a substantial fraction of misclassified predictions were within the most uncertain epochs [64]. Most recently, Rusanen et al. evaluated several softmax-based measures of aSAGA, a convolutional neural classifier, and reported effective identification of predictions in the mismatch to the consensus-scoring of 5 scorers [124].

The integration of sleep-scoring algorithms into clinical practice demands a deep understanding of the physician's real needs and expectations. However, these are seldom

considered in existing work, which approaches this problem in isolation from the human experts. Our study builds upon the U-Sleep algorithm, a state-of-the-art DL-based sleep scoring model trained on a broad spectrum of open-access clinical databases. Considering the intrinsic limitations of sleep scoring, rather than just aiming to improve the model's epoch-wise performance, which might already be at its ceiling level due to the inter-scorer variability, our study seeks to integrate this established system in a manner that actively involves physicians.

By investigating various strategies for pinpointing the least confident predictions and streamlining their review, we aim to redefine the collaboration between sleep-scoring algorithms and clinicians. Utilizing clinically rich Berner Sleep Data Base (BSDB) [125], we systematically investigate (i) *the optimal strategies to gather uncertain sleep stage predictions for the physicians' review* and based on that we (ii) q*uantify the volume of predictions that need to be reviewed (ie, physician's effort) to reach certain agreement benchmarks*. Leveraging details on physicians involved in scoring of individual BSDB PSGs, we robustly assess the efficacy of our combined system integrating the sleep-scoring algorithm with uncertainty estimation, considering both *in-domain* (ID), and potentially more challenging *out-of-domain* (OOD) test data.

Semi-automated approaches for sleep staging have been explored in various modalities and frameworks [48], [64], [93], [94], [122]–[124]. However, comprehensive testing of these methods against their limitations has been relatively sparse. To the best of our knowledge, our study is the first one extensively addressing a wide range of challenges specific to semi-automated scoring. This includes an in-depth examination and adaptation to individual scoring tastes of single (OOD) physicians, the impact of different sleep-disorder diagnoses on our approach's validity, the metrics employed, as well as the dimensions and diversity of the datasets involved.

2.2 Materials and Methods

2.2.1 Dataset

For our primary evaluations, we exploited the Berner Sleep Data Base (BSDB) from our partner clinic, Inselspital, University Hospital Bern. A total of 8,832 PSGs have been collected from 2000 to 2021 on individuals covering the whole spectrum of age (0–91 years), sleep disorders, as well as healthy controls. The signals were recorded at 200 Hz and, across 20 years of data collection, scored manually by more than 10 senior and 50 assistant physicians according to the AASM rules. To match older recordings scored according to Rechtschaffen and Kales with AASM standard, the N3 and N4 stages were merged into a single-stage N3. Secondary usage of the dataset was approved by the local ethics committee (KEK-Nr. 2020–01094). Participants provided written general consent upon its introduction at Inselspital in 2015, and data were maintained with confidentiality. Most individuals underwent PSG due to the suspicion of a sleep disorder. Together 66 individuals represented healthy subjects that took part as controls in clinical trials. The BSDB provides various levels of diagnoses based on individual tests (eg, actigraphy- or PSG-based). For our evaluations, we considered the clinically most relevant conclusive diagnoses made by physicians considering all test-based diagnoses, clinical anamnesis, and the context. The amount of available conclusive diagnoses is compared to the test-based ones smaller but provides the most reliable and highly trustworthy information.

For the purpose of our research, we divided the BSDB into three parts: one in-domain (ID) subset – consisting of training, validation, and test data splits consisting of PSGs, each scored by one of >50 physicians – used for optimization and baseline evaluation of the algorithmic approaches adopted – and, utilizing the information about the scorers, we created two out-of-domain (OOD) held-out subsets, each containing PSGs scored by a unique senior physician not presented in ID data with potentially different "scoring taste" than the population of ID-included physicians. Hence, such stratified evaluations on OOD subsets represent a more robust generalizability assessment close to the scenario happening in clinics, where typically a single physician takes decisions (e.g., about scoring, diagnosis). As one patient can have multiple PSGs recorded, all data splits were done per subject, assuring that

the individual's data are present only in one subset. A summary of data splits with respect to the number of PSGs, physicians involved, and demographic characteristics of subjects is provided in Table 2.1. Moreover, Table 2.2 provides details on the occurrence of different classes of sleep disorders among conclusive diagnoses of subjects.

In addition to BSDB, part of our work replicated the training of the sleep-scoring algorithm U-Sleep, using 19,578 PSGs from 13 open-access databases. A detailed description of these data, including demographic characteristics, is provided in the original publication [60].

Table 2.1: Demographic characteristics of BSDB subjects across data splits.

Domain	Scorers (N)	Split	PSG (N)	Age μ (σ) – median – min – max	Gender (%, ♂–♀)
ID	$>$ 8 SP, $>$ 50 AP	Train	4,245	49.22 (16.40) - 51 - 2 - 88	64.28–35.72
		Validation	226	52.66 (21.45) - 60 - 8 - 84	67.71–32.29
		Test	423	50.48 (20.32) - 55 - 2 - 86	65.57–34.43
OOD1	1 SP	Test	1,966	48.90 (18.60) - 52 - 0 - 91	64.65–35.35
OOD2	1 SP	Test	1,972	46.93 (20.06) - 50 - 0 - 86	60.92–39.08
TOTAL	$>$ 10 SP, $>$ 50 AP	–	**8,832**	**48.82 (18.25) - 51 - 0 - 91**	**63.76–36.24**

Abbreviations: μ, mean age per group; σ, standard deviation of age per group; min, minimum age; max, maximum age; %, percentage; SP, senior physician; AP, assistant physician.

Table 2.2: Occurrence of sleep disorder classes across BSDB conclusive diagnoses and data splits.

	Domain					
Diagnosis class	ID train	ID Validation	ID test	OOD1 test	OOD2 test	**ALL**
HE	27	2	3	12	22	66
INS	106 + 15	8 + 1	17 + 2	31 + 5	43 + 4	205 + 27
SDB	247 + 156	16 + 8	34 + 17	91 + 33	124 + 18	512 + 232
CDH	171 + 30	10 + 2	22 + 5	54 + 1	115 + 10	372 + 48
CRD	11 + 1	0 + 0	2 + 0	1 + 0	5 + 0	19 + 1
PSD	75 + 9	7 + 0	6 + 0	22 + 1	44 + 0	154 + 10
SMD	74 + 5	5 + 0	7 + 0	18 + 1	33 + 0	137 + 6
IS	227 + 11	13 + 1	26 + 1	77 + 1	127 + 1	470 + 15
DSS	26 + 0	1 + 0	2 + 0	7 + 0	16 + 0	52 + 0
Multiple disorders	418	26	52	128	205	829
Single disorders	227	12	25	42	33	339
Other or unknown	3573	186	343	1,784	1,712	7,598
TOTAL	4,245	226	423	1,966	1,972	8,832

Notes: Columns indicate individual data subsets: ID (training, validation, testing) and two OOD test sets (OOD1, OOD2), summing up to ALL. Rows indicate the number of subjects according to conclusive diagnoses class indicated by abbreviations described below. Row *Multiple disorders* indicates the number of subjects with multiple classes of sleep-disorders, *Single disorders* the number of subjects with a single sleep disorder, and *Other or unknown* the number of subjects with no or unknown *conclusive* diagnosis. TOTAL is equal to HE + Multiple disorders + Single disorder + Other or unknown. At the cell level of rows (INS to DSS), the sum refers to the number of subjects having multiple disorders including that given class plus the number of subjects having that specific class only.

Abbreviations: HE, healthy controls; INS, insomnia disorders; SDB, sleep-disordered breathing; CDH, central disorders of hypersomnolence; CRD, circadian rhythm sleep-wake disorders; PSD, parasomnia-related sleep disorders; SMD, sleep-related rhythmic movement disorders; IS, isolated symptoms and normal variants; DSS, findings specific to day-time sleep studies.

2.2.2 U-Sleep: The Sleep Scoring Algorithm

The U-Sleep, introduced by Perslev et al. [59], is a deep convolutional neural network for sleep stage classification inspired by the U-Net, an architecture originally used for image segmentation [126]. The U-Sleep takes as its input at least one pair of EEG-EOG channels (re)sampled at 128 Hz and outputs an array of softmax values quantifying the plausibility of each signal window (epoch) of a specified length, usually 30 seconds, to represent one of the 5 sleep stages. If more input channel-pairs are available, the U-Sleep averages the softmax outputs over all of them. The architecture of U-Sleep consists of an encoder-decoder part – compressing and decompressing the input signal using convolutional operations – followed by a classifier layer.

In-depth technical details on the U-Sleep architecture, including the preprocessing steps implemented to unify signals from different devices, and the training process, are thoroughly described in the original work [59]. This study also reports the state-of-the-art performance on 16 databases of more than 15,000 participants, achieving an average F1-score of 79%. The robustness of U-Sleep was confirmed even after its original implementation was corrected for a channel-derivation bug, achieving an average F1-score of 76.5% [60].

Our work replicated the training run on 13 open-access databases of 19,578 PSGs using the most recent implementation of U-Sleep [60]. Based on that, we exploit the rich BSDB and fine-tune (re-train) the U-Sleep using training and validation ID-splits as described in Table 2.1. Finally, we use such fine-tuned U-Sleep as a basis for the selection of the most suitable approach of uncertainty estimation to enable an efficient review of predicted hypnograms by physicians. The generalizability of both sleep scoring and predictive uncertainty-quantification approaches were rigorously evaluated on the ID test set and two single-scorer OOD subsets of the BSDB.

2.2.3 Estimation of Predictive Uncertainty

In advancing sleep scoring algorithms for clinical practice, one crucial component is the quantification of predictive uncertainty, which encompasses both epistemic and aleatoric aspects. Epistemic uncertainty, in a sleep-scoring context, arises from the variability in how physicians interpret AASM guidelines, leading to $\sim$20% noise in sleep-stage labels due to $\sim$80% inter-scorer agreement. On the other hand, aleatoric uncertainty, inherent in the variability of sleep patterns themselves, represents a natural randomness that cannot be mitigated. In this section, we elaborate on our approach with the U-Sleep classifier. First, we detail measures of predictive uncertainty based on the classifier's softmax output. Next, we describe adapting an auxiliary confidence network, specifically designed for sleep-related time-series representations derived from the U-Sleep, to estimate confidence in its predictions. The terms uncertainty and confidence can be understood as complementary and will be used according to the appropriateness of the context. The integration of uncertainty quantification is pivotal not only in elevating the trustworthiness of the automated sleep-scoring solutions but also in enabling physicians to efficiently review and verify algorithm-proposed predictions.

Softmax-Based Measures

The confidence level of a classifier's predictions can be gauged from its softmax output, which can be graphically represented as *hypnodensity* [61]. This can be analyzed either visually or, when uncertain epochs should be automatically gathered, by numerical assessment of the softmax values. At its simplest, the maximum value of the predicted softmax can be perceived as a representation of the epoch's likelihood of belonging to a specific class (i.e., sleep-stage). The closer the max-softmax is to 1, the higher the confidence, while lower values indicate uncertainty. There are a variety of measures, rooted in softmax outputs, that can be employed to discern these uncertainties. For instance, several works employed entropy-based measures because as entropy rises, the distribution of softmax values becomes more uniform [64], [122], [124].

Regardless of the chosen measure, uncertain predictions from each predicted hypnogram can be highlighted in two ways: (i) *by showcasing a fixed percentage of the most uncertain*

epochs or (ii) *by indicating epochs that surpass a specific value threshold*. The latter is more advocated as it may consider the sampling distribution of classification accuracy. Moreover, the fixed-percentage approach has greater potential to introduce undesired results (false positives/negatives) if the predetermined percentage does not coincide with the actual amount of misclassified epochs. In our research, we sought methods that adeptly identify uncertain predictions for subsequent review by clinical experts. A comprehensive mathematical detailing of all measures employed in our work is provided in Table 2.3, whereas a comparison in terms of their ability to discern predictions discordant with human scoring is presented in Results.

Table 2.3: Measures of prediction uncertainty based on U-Sleep softmax output.

	Measure	Notation	Mathematical formula
I	Average softmax-entropy	$\bar{p}_{\text{entr}}$	$-\frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{5} p_{mk}\log_2 p_{mk}$
II	Average softmax-ratio	$\bar{\rho}$	$\frac{1}{M}\sum_{m=1}^{M}\frac{1}{5}\sum_{k=1}^{5}\frac{p_{mk}}{\max(p_m)}$
III	Average softmax-standard-deviation	$\bar{\sigma}$	$\frac{1}{M}\sum_{m=1}^{M} SD(p_m)$
IV	Maximum of majority-softmax	μ	$\max\left(\frac{1}{M}\sum_{m=1}^{M} p_m\right)$
V	Standard deviation of majority-softmax	σ	$SD\left(\frac{1}{M}\sum_{m=1}^{M} p_m\right)$
VI	Fixed % according to μ	$\mu\%$	–
VII	Fixed % according to σ	$\sigma\%$	–

Notes: Uncertainty measures adapted for majority-voting mechanism of the U-Sleep classifier
Abbreviations: M, total number of input channel-pairs used; m, index over M; k, index over 5 classes (i.e., sleep stages); p_{mk}, probability (i.e., softmax-value) of the k-th class based on the m-th input channel pair; p_m, probability vector (i.e., softmax) of 5 classes based on the m-th input channel pair; max, maximum; SD, standard deviation.

Uncertainty Quantification Using an Auxiliary Confidence Network

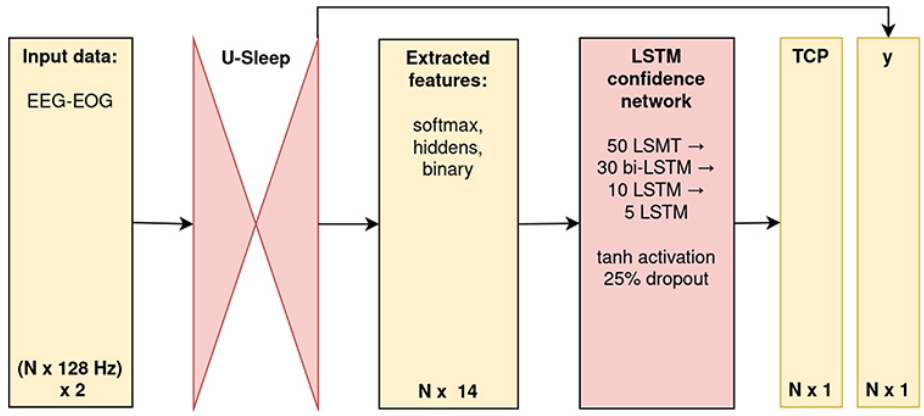
Neural networks, while powerful, often exhibit overconfidence, manifested as a disparity between the predicted softmax value and the actual probability of an observation belonging to a specific class [127]. This may limit the use of softmax-base measures to gather uncertain predictions accurately. To counteract this issue, Corbiere et al. proposed an auxiliary confidence network, which aims to estimate the *True Class Probability* (TCP) score, designed to work in tandem with an already-trained classifier network [128]. The TCP is defined as the value of the predicted softmax that aligns with the true label, meaning, for misclassified predictions, it diverges from the softmax maximum value. Upon the completion of classifier training, the TCP scores are extracted from training and validation data and serve as a target for the confidence network. This positions the training of the confidence network as a regression problem, where the objective is to predict the TCP – a single float value within the (0, 1) range – for each observation. In the original work, the confidence network was applied to image data, supplementing a convolutional network classifier, which involved reusing the classifier's architecture and its pre-trained weights, adding additional layers to facilitate the prediction of the TCP outcome, and finally optimizing the modified architecture [128].

Our contribution extends this idea specifically to PSG time-series data. Leveraging the U-Sleep output, we designed a lightweight sequence-to-sequence long-short-term-memory (LSTM) confidence network [129]. For each EEG-EOG input channel-pair of U-Sleep, our confidence network is fed by representations extracted from U-Sleep layers, including the 5-dimensional softmax output, the binary code of the same dimensionality as softmax indicating the predicted class, and the five-dimensional hidden features extracted from the layer preceding the softmax. The adoption of a bidirectional-LSTM-based architecture was driven by our beliefs that the uncertainty in predicting sleep stages is intrinsically tied to sequential information – namely, the representations preceding and succeeding a given epoch. Recognizing the functional dependencies in the softmax output (that sums up to 1), we applied to it the additive log-odds ratio (ALR) transformation, which reduces the dimension by one (i.e., to 4) and decreases the co-linearity [130]. Building on the premise that combined data offers a richer perspective for identifying the most uncertain predictions, we fed the confidence network with all such extracted features simultaneously. The final architecture of our

confidence network had 35,628 parameters and consisted of three main parts: an input layer with batch normalization; 4 hidden layers (LSTM of 50 neurons, bidirectional-LSTM of 30 neurons enabling information flow from the past as well as future states, two LSTMs of 10 and 5 neurons) returning sequences, with tanh activation function and 25% drop-out; and a final layer with an output LSTM neuron with the custom activation function, (tanh(x)+1)/2, returning a sequence in desired (0, 1) range, corresponding to the predicted sequence of TCP confidence scores for each PSG-epoch. These are then, consistent with U-Sleep's mechanism, averaged across all input channel-pairs used. The design of our LSTM-based confidence network is presented in Figure 2.1

The TCP confidence score using a more complex input information processed by a specifically designed neural network extended the set of rather simpler softmax-based measures. Our evaluations focused on their in-depth comparison in terms of identifying U-Sleep-predicted epochs that do not align to the physician's scoring, forming a basis for creating a system that allows physicians to effectively utilize automatic sleep scoring algorithms.

Figure 2.1: Schematic overview of the implemented pipeline.



Notes: An EEG-EOG channel-pair is used as an input for the U-Sleep classifier. Using the trained U-Sleep, several representations are extracted (softmax; binary code indexing the predicted class; hidden representations - hiddens - from the layer preceding softmax) and used as an input for the confidence network evaluating the True Class Probability (TCP) confidence score. The hypnogram predicted by U-Sleep (y) is provided jointly with the assessment of predictive uncertainty (1-TCP) to guide an efficient review by a physician.

2.2.4 Utilizing Uncertainty Estimates for an Efficient Review of Predicted Hypnograms

Our analysis, tailored towards the efficient use of uncertainty estimates for the review of predicted hypnograms, was guided by a three-tiered evaluation approach: (i) *selection of the best-suited uncertainty measure*; (ii) *statistical evaluations of its discriminative power*; and (iii) *the impact-evaluation when physicians rescore the most uncertain predictions gathered*. While the first two aspects focus on the technical aspects, the conclusive part evaluates the practical implications, comparing the physician's effort – quantified as the amount of epochs reviewed – in relation to the boost of the agreement between their scoring taste and partially reviewed predictions of the scoring algorithm.

Best-Suited Uncertainty Measure

Initially, in order to pinpoint the most suitable uncertainty measure, we treated identifying epochs diverging from human scoring as a binary classification task. The diverging epochs from human scoring, ie, the U-Sleep-misclassified predictions, were considered a positive class. Using this setup, we selected the most apt measure based on their Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curve performances. The choice of ROC and PR curves stems from their ability to handle class imbalances and effectively comparing the true-positive against false-positive rates.

Statistical Tests to Assess the Discriminative Power of the Superior Uncertainty Metric

Upon identifying the superior metric, we further sought to statistically assess its efficacy in two distinct manners. Firstly, we proposed the null hypothesis H_{01}: *"There is no significant difference between the on-subject mean-aggregated uncertainty scores of epochs congruent with human scoring and those diverging from it."* In other words, this would imply that the uncertainty in correctly scored epochs would be the same as for the misclassified ones. With H_{01}, we aimed to test whether predictions in line with human scoring systematically differed from those diverging in terms of their uncertainty score, effectively probing the metric's ability to distinguish between correctly versus incorrectly classified epochs.

Further, the null hypothesis H_{02} postulated: *"There is no significant correlation between the mean-aggregated on-subject uncertainty scores and the on-subject classification performance metrics."* In other words, that would imply that, e.g., classification accuracy is not associated with uncertainty levels. The H_{02} aimed to assess the relationship between the uncertainty attached to predictions and the classification performance on a per-subject basis.

Both assessments were conducted separately for ID and OOD data, with consideration of sleep-disorder status of individuals. Given the skewed non-normal nature of the uncertainty measures with bounded value ranges, the non-parametric bootstrap was employed to calculate confidence intervals (CI) to assess both hypotheses [131].

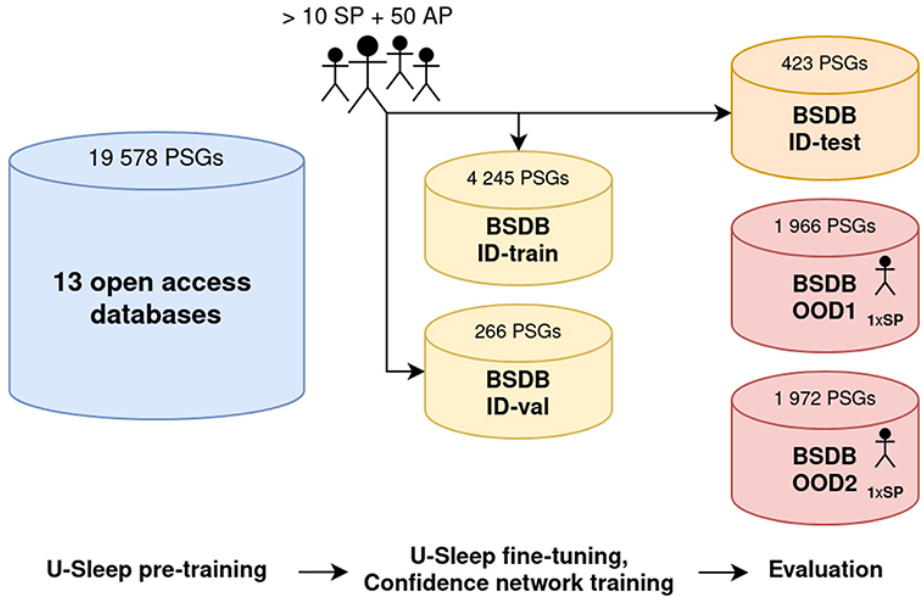
Impact-Evaluation of Physician Intervention on Uncertain Epochs

The culmination of our analysis revolved around varying the threshold employed to discern the uncertain epochs for the superior uncertainty metric identified. Under each threshold specification from a predefined grid, a physician review was enacted, with discordant predictions being rectified and agreeing epochs being kept. Subsequently, the classification metrics were recalculated to encapsulate this simulated physician's intervention. While the relation between increased reviewed epochs and monotonic performance improvement is evident, our objective was to quantify the rescoring effort required to meet distinct performance benchmarks. This examination was undertaken across both ID and OOD test data splits, fortifying the robustness of our conclusions. Further, in order to make fair comparisons with existing research, we enumerated the performance improvements across diverse metrics: accuracy (Acc), weighted F1-score ($F1_w$), and Cohen's kappa (κ).

2.3 Results

In this section, we provide the main findings with respect to the algorithmic methods exploited and developed (U-Sleep algorithm along with the auxiliary confidence neural network), and their validation on individual data domains, as depicted within the workflow in Figure 2.2.

Figure 2.2: Schematic overview of the datasets, their sizes, and purposes.



Notes: A set of 13 open-access datasets (in blue) was used for the baseline training of the U-Sleep. The middle and right parts of the schema relate to the evaluations on BSDB. Its ID part refers to PSGs each scored by one of more than 50 assistants and 10 senior physicians. The ID training and validation splits (in yellow) were used to fine-tune U-Sleep and, subsequently, to train the confidence network. Baseline evaluation of both algorithmic approaches was performed on the ID-test data (in orange). Their robustness was further evaluated on two OOD test sets (in red), each containing PSGs scored by a unique SP.

2.3.1 U-Sleep Classification Performance

As a sleep scoring classifier, we employed U-Sleep and replicated the training experiment of its most recent implementation using 13 open-access databases of 19,578 PSGs [60]. Next, the model was fine-tuned on the BSDB, leveraging the ID training and validation splits as elaborated in Table 2.1. The U-Sleep optimization based on minimization of the categorical cross-entropy loss converged after 539 training epochs. To ensure a comprehensive comparison with existing research, we enumerated three distinct classification performance metrics: Acc, $F1_w$, and κ, computed in three different ways: epoch-wise (pertaining to all 30-second windows in the relevant data split), as well as subject-wise mean- and median-aggregated. Table 2.4 summarizes the performance across the ID and the two OOD test data. The results indicate that the epoch-wise performance on ID (test) slightly exceeded that of the OOD2 and was marginally inferior to OOD1, with a maximum difference of 2.9% in the $F1_w$ between ID vs OOD1. These findings were consistent for on-subject metrics. Noteworthy, on the ID test split, which contains "scoring tastes" of more than 50 different physicians involved in scoring of PSGs, U-Sleep reached the subject-wise agreement level of κ = 76.2% that corresponds to the interscorer agreement of κ = 76% reported in the literature [15], [65], [88]. This points to the robustness of U-Sleep's scoring ability in line with the theoretically justifiable performance ceiling that can be achieved on human-scored hypnograms. Marginal over- and under-performance on OOD data splits can be attributed to the greater or lesser consistency of the given (split-specific) senior physician with the "overall" population scoring pattern encoded in U-Sleep.

Table 2.4: Classification performance of U-Sleep across individual data splits.

Domain	Metric	Epoch-wise	Subject-wise mean	Subject-wise median
ID-test	Acc	82.5	82.1	84.5
	$F1_w$	82.8	82.4	85.3
	κ	75.0	71.2	76.2
OOD1	Acc	84.2	84.5	86.4
	$F1_w$	85.0	85.5	87.4
	κ	77.6	76.0	78.8
OOD2	Acc	80.7	80.8	82.7
	$F1_w$	80.5	81.4	83.4
	κ	73.3	71.1	73.8

Notes: Epoch-wise performance calculated over all 30-second windows present in individual data splits. Mean and median subject-wise metrics are calculated as performance achieved on individual-specific hypnograms.

2.3.2 Evaluation of Approaches for Uncertainty Estimation

The primary objective in this phase was to pinpoint the best approach that adeptly identifies U-Sleep-predicted epochs that deviate from human scoring. This consisted of two main strands of investigation: comparing softmax-based uncertainty metrics and evaluating the confidence scores based on the adapted confidence neural network.

Softmax-Based Measures

We initially took into consideration all the softmax-based metrics, as delineated in Table 2.3. The metrics (I–V) identify uncertain epochs based on a distributional threshold, while metrics (VI, VII) are designed to accumulate a predetermined percentage of the most uncertain predictions. The fixed-percentage strategies do not include an approach based on the softmax ratio ($\bar{\rho}$) as it is monotonically dependent on the maximum of the softmax (μ) and would lead to the same results. Calculation of these metrics was straightforward, as they involved only the U-Sleep softmax output based on each input channel-pair. The performance of individual measures in terms of identifying predictions discordant from human scoring is listed in Table 2.5. The majority of the metrics achieved comparable results with the superiority of the distributional-threshold-based metrics over the fixed-percentage strategies, confirming the need for a flexible approach adapting to possibly different amounts of difficult-to-score (uncertain) epochs per PSG. The best performing approach was μ – the maximum of the majority-softmax (= softmax averaged over all input channel pairs) – reaching AUROC of 76.5% on the ID-test and 82.4–81.1% on the two OOD sets.

Table 2.5: Performance of uncertainty measures in detecting U-Sleep predictions that deviate from human scoring across data splits.

Domain	Evaluation metric	$\bar{p}_{\text{entr}}$	$\bar{\rho}$	$\bar{\sigma}$	μ	σ	μ%	σ%	TCP*
ID-test	AUROC	76.4	75.7	76.2	76.5	64.3	59.1	56.5	**85.7***
	AUPR	39.7	41.3	41.0	42.9	30.2	36.5	31.4	**63.1***
OOD1	AUROC	80.1	82.0	81.6	82.4	75.4	60.6	57.2	**85.6***
	AUPR	38.8	42.0	41.0	43.5	41.0	33.3	26.8	**53.6***
OOD2	AUROC	79.6	80.8	80.6	81.1	75.0	59.9	57.1	**82.5***
	AUPR	43.2	45.0	44.6	45.8	34.1	36.9	31.4	**50.7***

Notes: Performance assessment as the % of the AUROC and AUPR curves for the softmax-based measures from Table 2.3 and the True Class Probability (TCP) score based on confidence network. Bold font highlights the best performance obtained among compared metrics.

Auxiliary Confidence Network

Our evaluations continued with the auxiliary confidence network leveraging the joint information of the transformed softmax output and the hidden representations extracted from U-Sleep to predict the True Class Probability (TCP) score. We trained the confidence network on the ID training and validation splits, targeting the actual TCP scores calculated based on predictions of the already trained U-Sleep classifier. The training was based on minimizing the mean-absolute-error (MAE) loss, adopting mini-batches of U-Sleep-derived features for one PSG channel pair (EEG-EOG) at the time, and adhering to the default configurations of the Adam optimizer in Tensorflow 2.6.0. The training process achieved convergence after 16 epochs, marking a validation MAE of 0.0827. This indicates the confidence network's capability to predict the TCP with an average error of 8.27% in probabilistic terms. It is worth noting that the training set incorporated epochs labeled as "unknown" by physicians, reflecting the inherent challenges in scoring such signals, often due to untouched electrodes yielding constant (zero) signal. These particular epochs were assigned a target TCP of 0, given that none of the softmax values would match the correct class (i.e., sleep-stage).

Having the trained confidence network, we evaluated how its predicted TCP-score performs to detect discordant epochs. Focusing on the last column of Table 5, we observe its superiority in comparison to all simpler softmax-based approaches across all test data subsets. It outperformed the other approaches in terms of both ROC and PR assessments, reaching AUROC of 85.7% on ID, 85.6–82.5% on the two OOD sets, and AUPR of 63.1% for ID and 52.3–50.7%, respectively. Furthermore, the robustness of the confidence network was confirmed, as it delivered comparable performance on both ID and OOD splits, highlighting its generalizability to potentially different scoring patterns introduced by different senior physicians. Given its demonstrated efficacy, the TCP confidence score was selected as the key metric for the following evaluations simulating physician's interventions, focusing on the review and eventual correction of the most uncertain predictions.

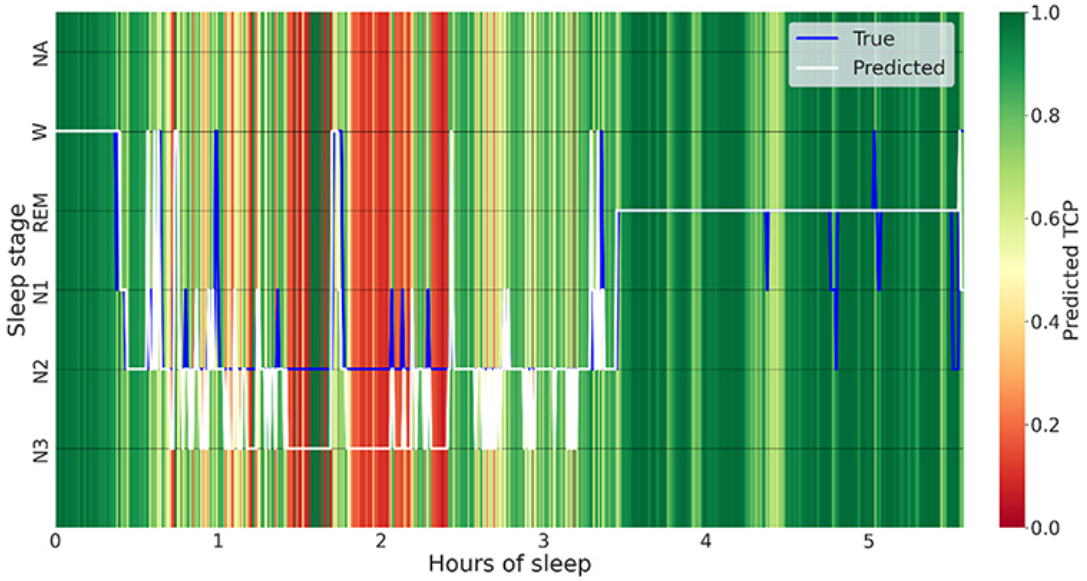
Confidence-Supplemented Hypnogram

Using the TCP as the most reliable uncertainty quantification measure, Figure 2.3 depicts the combined output of the U-Sleep-predicted hypnogram (in white) with the estimated confidence TCP-scores as a green-red color scale in the background. This dual output is a result of our final pipeline, depicted as a diagram in Figure 3, detailing the process of transforming original biosignals into a joint presentation of predicted sleep stages and their associated confidence levels. Such visual representation is designed to guide the physician in identifying specific segments of the PSG that deserve closer review. For demonstration, the actual physician's scoring on given PSG, referred to as true, is depicted in blue. A close examination reveals that segments with lower predicted TCP scores often (e.g., 1:30–2:30 h of sleep) predominantly align with U-Sleep misclassifications. In contrast, regions with higher scores (e.g., from 3:30 h onwards) mostly point to accurately scored epochs. It is important to note that since the estimated TCP scores are model-derived, occasional discrepancies can arise. For instance, around 1:45 h, a brief period marked with high confidence corresponds to discordant scoring. Even though this segment erroneously indicates high confidence, its neighborhood areas of low confidence might draw physician's attention for a review. Despite occasional inconsistencies, the results from Table 2.5 indicate that TCP-score has the best ability to identify discordant epochs.

2.3.3 Statistical Tests of on-Subject TCP Scores with Respect to Clinical Diagnosis

Further, we investigated in-depth the discriminative power of the TCP-score to reveal discordant predictions. Firstly, to evaluate H_{01}, we calculated the on-subject difference between averaged TCP-scores of predictions that align and those that disagree with human scoring: $d_i = \overline{TCP}_{i,correct} - \overline{TCP}_{i,incorrect}$. Next, for the evaluation of H_{02}, the on-subject performance metrics (Acc_i, $F1_{w,i}$, κ_i) and the overall average TCP score ($\overline{TCP}_i$), for each subject's predicted hypnogram were calculated. The $\overline{TCP}_i$ can be understood as an assessment of the

Figure 2.3: Combined output of predicted and physician-scored hypnograms with associated confidence scores.



Notes: Combined output of the predicted hypnogram (in white), associated TPC confidence scores (in the background), and physician-scored hypnogram (in blue), for a 44-year-old female diagnosed with hypersomnolence. On-subject (Acc, $F1_w$, κ) of (79.2, 72.2, 61.5)%, respectively. On-subject average TCP of 0.74. For correctly and incorrectly classified epochs, the average on-subject TCP was 0.87 and 0.41, respectively.

confidence over the entire predicted hypnogram of a given subject. We employed a non-parametric bootstrap approach, with 5000 repetitions, for both hypotheses to compute 95% confidence intervals (CIs). Having a database rich in sleep-disorder diagnoses enabled us to assess both hypotheses considering individual classes of diagnoses, as described in Table 2.2. To assess the generalizability of our findings, we considered subjects from the ID-test and the two OOD test data with confirmed conclusive diagnoses. Since the subjects – except for healthy controls – suffer in many cases from several sleep disorders, we always included in a given class all who have at least one corresponding diagnosis. Both hypotheses were assessed on disorder classes of at least 10 subjects, separately on the ID test data, and – to achieve a larger sample size in each class – the pooled OOD data.

Table 2.6 gives an overview of bootstrapped 95% CIs and the medians related to H_{01} for each diagnosis class considered. Based on the CIs obtained, H_{01} can be rejected (p-value < 0.05 in all cases), and one can conclude that the difference between the mean-aggregated TCP-scores of aligning and discordant predictions significantly differs and is consistently greater than 0. All that across the entire diagnosis spectrum, on both ID and OOD test domains. The median differences ranged as 0.20–0.23 and 0.19–0.26, for ID and OOD, respectively, which affirms that the TCP-score was in terms of a probability about 20% lower for the discordant predictions. In an extension of our analysis, we conducted the same evaluation on a subgroup of 76 children under 6 years old, using pooled OOD data. Compared to the mean classification metrics presented in Table 2.4, U-Sleep demonstrated lower scoring performance with Acc of 71.28%, $F1_w$ of 73.15%, and κ of 59.19%. This performance drop is likely attributable to specific AASM scoring rules applied to children. Nonetheless, the average on-subject difference between aligning and discordant TCP scores was significantly greater than zero, indicating a mean difference of 0.19 with a 95% CI of (0.17, 0.22). These findings suggest that the confidence network and the resulting TCP score can efficiently guide physicians on hypnogram and respective PSG sections needing review and potential correction, regardless of subject's diagnosis status, including pediatric cases.

Table 2.6: Bootstrap confidence intervals (CI) for differences in subject-level mean TCP scores between aligning and discordant predictions.

Domain:	ID-test			Pooled OOD		
Diagnosis Class	**Median**	**95% CI**	**N**	**Median**	**95% CI**	**N**
HE	–	–	3	0.26	(0.22, 0.30)	34
INS	0.21	(0.18, 0.24)	19	0.23	(0.21, 0.25)	83
SDB	0.20	(0.17, 0.22)	51	0.21	(0.20, 0.22)	266
CDH	0.23	(0.20, 0.27)	27	0.23	(0.21, 0.24)	180
PSD	–	–	6	0.19	(0.16, 0.21)	67
SMD	–	–	7	0.20	(0.18, 0.23)	52
IS	0.21	(0.18, 0.25)	27	0.22	(0.21, 0.23)	206
DSS	–	–	2	0.25	(0.21, 0.29)	23

Notes: Evaluations on ID-test data and pooled OOD data. Bootstrapped median stands for the estimate of the mean-difference, and the corresponding 95% CI are calculated as 2.5% and 97.5% quantiles of bootstrap resamples.
Abbreviations: HE, healthy controls; INS, insomnia disorders; SDB, sleep-disordered breathing; CDH, central disorders of hypersomnolence; CRD, circadian rhythm sleep-wake disorders; PSD, parasomnia-related sleep disorders; SMD, sleep-related rhythmic movement disorders; IS, isolated symptoms and normal variants; DSS, findings specific to day-time sleep studies.

Further, Table 2.7 relates to H_{02} and details the bootstrapped 95% CIs for the correlation between the average on-patient TCP score and the classification performance metrics. Based on the CIs obtained, we conclude that for all diagnoses of both ID and OOD test data, the correlation with any performance metric was consistently significant (p-value < 0.05 in all cases) and positive. The TCP correlated – on average – the most with the accuracy with a range of 0.67–0.74 across individual diagnosis classes of ID test data, and of 0.58–0.81 for OOD data. Consistent findings were identified even for the 76 OOD children aged under 6 years, where TCP was significantly positively correlated with all the performance metrics: 0.62 with 95% CI of (0.43, 0.76) for Acc, 0.56 (0.36, 0.72) for $F1_w$, and 0.60 (0.41, 0.75) for κ. These findings suggest that the aggregated TCP score can efficiently pinpoint subjects whose biosignals are challenging to classify and also those with high prediction performance, including children with different AASM scoring rules applied.

Table 2.7: Bootstrap confidence intervals (CI) for correlations between subject-level mean TCP scores and performance metrics.

Diagnosis Class	Performance Metric	ID-test		Pooled OOD	
		Median	95% CI	Median	95% CI
HE	Acc			0.74	(0.59, 0.91)
	κ			0.67	(0.50, 0.89)
	$F1_w$			0.60	(0.41, 0.84)
INS	Acc	0.67	(0.46, 0.91)	0.58	(0.46, 0.78)
	κ	0.56	(0.28, 0.88)	0.59	(0.47, 0.84)
	$F1_w$	0.63	(0.39, 0.91)	0.49	(0.36, 0.72)
SDB	Acc	0.71	(0.62, 0.85)	0.71	(0.66, 0.80)
	κ	0.69	(0.58, 0.85)	0.68	(0.62, 0.77)
	$F1_w$	0.57	(0.44, 0.78)	0.65	(0.59, 0.75)
CDH	Acc	0.72	(0.55, 0.90)	0.75	(0.69, 0.84)
	κ	0.64	(0.45, 0.86)	0.72	(0.66, 0.82)
	$F1_w$	0.58	(0.35, 0.85)	0.68	(0.61, 0.79)
PSD	Acc			0.81	(0.74, 0.89)
	κ			0.81	(0.74, 0.90)
	$F1_w$			0.78	(0.70, 0.87)
SMD	Acc			0.63	(0.47, 0.84)
	κ			0.54	(0.37, 0.79)
	$F1_w$			0.55	(0.39, 0.79)
IS	Acc	0.74	(0.58, 0.90)	0.70	(0.64, 0.80)
	κ	0.74	(0.59, 0.90)	0.64	(0.57, 0.76)
	$F1_w$	0.62	(0.40, 0.87)	0.63	(0.56, 0.75)
DSS	Acc			0.62	(0.34, 0.92)
	κ			0.62	(0.36, 0.93)
	$F1_w$			0.49	(0.21, 0.86)

Notes: Evaluations on ID-test data and pooled OOD data. Bootstrappped median stands for the estimate of correlation with a performance metric, and the corresponding 95% CI are calculated as 2.5% and 97.5% quantiles of bootstrap resamples.
Abbreviations: HE, healthy controls; INS, insomnia disorders; SDB, sleep-disordered breathing; CDH, central disorders of hypersomnolence; PSD, parasomnia-related sleep disorders; SMD, sleep-related rhythmic movement disorders; IS, isolated symptoms and normal variants; DSS, findings specific to day-time sleep studies.

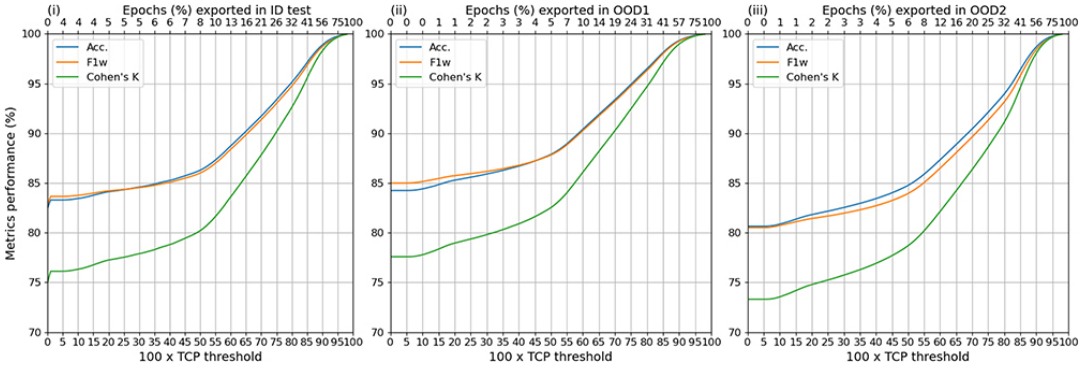
2.3.4 Performance Boost Under Physician's Intervention

In the final part of our evaluations, we aimed to quantify the potential improvement in sleep-scoring classification performance when the most uncertain predictions underwent physician's review. We simulated an intervention in which predictions with a TCP confidence score falling below a designated threshold, incremented in 0.01 steps across the [0,1] range, were set aside for human assessment. Within this set, predictions that did not align with the physician's assessment were subsequently adjusted to reflect the physician's scoring evaluation. Alongside observing the uplift in performance, we also monitored the amount of predictions subjected to review. This amount is indicative of the physician's time spent on re-scoring, prompting us to quantify the effort needed to reach specific performance benchmarks.

Figure 2.4 depicts the impact of the physician's review on the classification performance for the ID-test and the two OOD test data. The lower x-axis depicts the TCP-score threshold used to gather uncertain predictions, whereas the upper x-axis to the corresponding total % of the epochs re-scored (ie, the physician's effort). The % refers to the aggregate over all PSGs in a given data split, as from each PSG were extracted only epochs below a given threshold and so, the individual % differed. At a TCP-threshold of 0, when no uncertain epochs are extracted, the performance as depicted on the vertical axis corresponds to the original epoch-wise performance as shown in Table 2.4. From Figure 2.4, we can observe a monotonic improvement in all the performance metrics with the increasing amount of epochs gathered for the review. Based on that, we can identify, that to reach, eg, at least 90% in all the evaluation metrics, a rescoring effort of about 26% for ID-test, 19% for OOD1,

and 27% for OOD2 is needed, respectively, whereas the corresponding TCP threshold lies consistently around 0.75.

Figure 2.4: Performance boost with physician review of epochs with TCP scores below a given threshold.



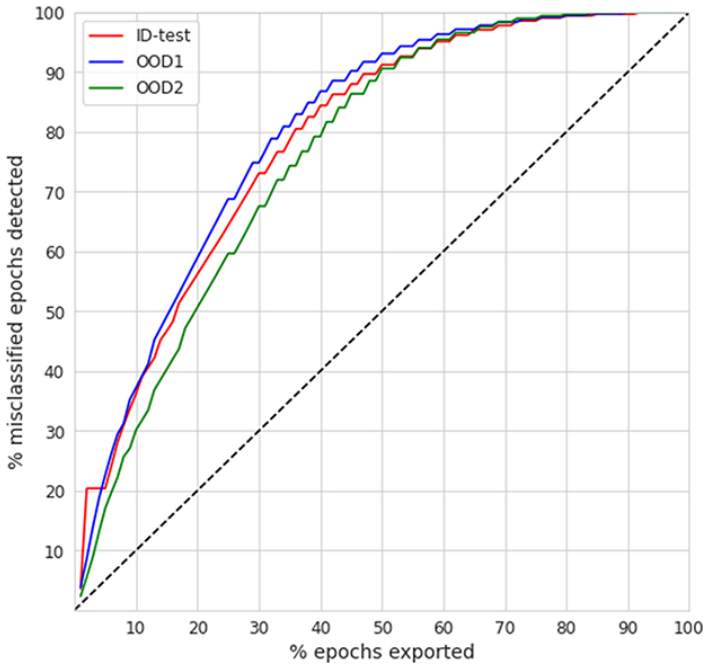
Further, based on Figure 2.4 and Table 2.8 summarizes the % of epochs needed to be reviewed to achieve the performance benchmarks of at least (80, 85, 90, 95)% for each evaluation metric, which we use for the comparison with other existing works in the Discussion. For example, to reach at least 90% in κ, a physician's review of 25.6% of epochs is needed on the ID-test, and 18.8–29.0% on the two OOD datasets.

Table 2.8: Rescoring amounts required to achieve target levels of sleep-scoring performance.

Domain	Metric	Desired performance level			
		80%	85%	90%	95%
ID-test	κ	7.6	15.7	25.6	41.5
	Acc	0	6.0	16.5	32.2
	$F1_w$	0	6.2	17.2	33.7
OOD1	κ	2.3	9.1	18.8	32.9
	Acc	0	1.1	9.8	25.5
	$F1_w$	0	0.1	10.5	25.5
OOD2	κ	8.3	17.3	29.0	44.0
	Acc	0	6.7	19.0	37.0
	$F1_w$	0	8.3	21.9	39.1

Finally, Figure 2.5 compares the rescoring effort based on an appropriate TCP-threshold in comparison to the % of all the misclassified epochs detected (ie, the true positive rate) per individual test data splits. The diagonal depicts a "random strategy", where physician's review would be conducted without any prior guidance on uncertain epochs. We observe that independently of the data domain, less than 50% of epochs need to be reviewed in order to detect at least 90% of all misclassified epochs. Similarly, to detect more than 95% of all misclassified epochs, a review of less than 60% of all epochs is needed. At a hypothetical 20% error rate, the 50% review effort with a corresponding detection of 90% out of all the discordant predictions leads to a boost of 18% resulting in a scoring performance of 98%, conforming with proposed clinical standards and being far beyond acceptable scoring error rates [121]. Since in our case is the error rate less than 20% for all domains (accuracy is always >80%) (as indicated in Table 2.4), the 50% review effort corresponds to obtaining almost perfectly aligned hypnograms with agreement above 98%

Figure 2.5: Review amounts (% of epochs exported) versus the % of discordant predictions gathered.



2.4 Discussion

Our study was motivated by a key clinical application in the field of sleep medicine, where physicians reach a consensus of about 76% when scoring PSG into sleep stages [15], [65], [88]. This level of agreement sets a technical limit on the accuracy metrics attainable when training scoring algorithms on multiple domains (scorers/databases). Consequently, when incorporating a scoring algorithm into clinical practice, its predictions must be subjected to a rigorous review by a human expert. If this is not guided to the uncertain regions of the predicted hypnogram and the respective PSG biosignals, such review may require a similar time effort as manual scoring done from scratch. Motivated by these challenges, we designed a pipeline where a state-of-the-art scoring algorithm is combined with an uncertainty estimation to guide the human review of the predicted hypnograms, with a particular focus on the quantification of the effort required to achieve certain performance benchmarks. We took advantage of the rich clinical database (BSDB) and evaluated our approach on both in-domain (ID) and the two out-of-domain (OOD) test data, considering individuals' conclusive sleep-disorder diagnoses. Such stratified analysis subjected our pipeline to a dual robustness test. In the case of the ID data, counting PSGs scored by >50 physicians, the evaluations related to the expected generalizability on an "average" pattern of sleep-scoring based on a broad population of physicians involved. On the other hand, the evaluations on OOD single-scorer splits were essential, because they assessed how well our system adapts to a real clinical setting, where PSG-scoring is performed by a single expert with a unique interpretation of the AASM rules.

As a sleep scoring classifier within our pipeline, we exploited the well-established U-Sleep, which we trained on 13 open-access databases and fine-tuned on ID (training and validation) data of BSDB. Such trained U-Sleep reached a robust performance of κ = 76.2% for ID test data and κ = (78.8, 73.8)% on the two single-scorer OOD sets, respectively.

Following that, we extensively investigated different uncertainty estimation approaches and assessed their performance on both ID and OOD datasets. Remarkably, our designed confidence network, specifically trained for PSG time-series data working in tandem with the U-Sleep, emerged as the superior approach, adeptly identifying predictions discordant with human scoring across both ID (AUROC = 85.7%) and the two OOD test data (AUROC of 85.6–82.5%). Identifying an approach that accurately pinpoints disagreeing predictions was a key prerequisite to enabling efficient review of predicted hypnograms by physicians.

Furthermore, our research extended into statistical examinations of the predicted uncertainty estimates, namely confidence scores based on our auxiliary network, leading to two pivotal conclusions: (i) the on-subject confidence scores were significantly different and lower for epochs discordant with human scoring, and (ii) the on-subject aggregated confidence scores significantly and positively correlated with all on-subject classification performance metrics. Both findings were consistent over the entire spectrum of sleep diagnoses present in both ID and OOD test data. Additional evaluations confirmed these conclusions even on 76 OOD children under 6 years of age, highlighting the generalizability of the predicted confidence scores for subjects with slightly different AASM scoring rules applied. These insights not only validate the efficacy of our approach for physician's review but also highlight its capacity to pinpoint sections of PSG biosignals that are inherently challenging to score, independently of the subject's diagnosis status, including pediatric cases.

As a pivotal component of our evaluations, we examined the extent to which guiding physicians in reviewing uncertain epochs could augment the efficacy of sleep staging. To attain a commendable classification performance of at least 90% in (κ, Acc, $F1_w$) metrics, our approach necessitated physicians to examine under 25.6% for κ, 16.5% for Acc, and 17.2% for $F1_w$ of the epochs on ID test data. For both OOD data, these figures were less than 29.0%, 19.0%, and 21.9%, respectively. These outpace the findings by Hong et al. [94] where about 35% and 25% of epochs needed a review to achieve a similar 90% rate in (κ, $F1_w$) on ID data primarily from sleep-disordered subjects. In the broader context, the review effort of our approach closely mirrors that of Phan et al. [64]. In their study on the Sleep-EDF dataset of healthy subjects, they reported a requirement to review 50% of epochs to identify 90% of all misclassified epochs. In our setup, with a dataset predominantly featuring sleep-disordered subjects, our efforts resonated closely, demanding a review of 45–50% of epochs, on both ID and OOD test data. Notably, the review of 50% of all the epochs leads, in our case, to an agreement of >98% for all ID and OOD test datasets. Furthermore, aiming for a more stringent identification of 95% of all misclassifications, our approach stands out, demanding a review of less than 60% of epochs on both ID and OOD test data - a subtle improvement over the 61.4% reported by Hong et al. [94]. In addition, our efforts are in line with the findings of the most recent work of Rusanen et al. [124], who identified about 90% of all misclassified cases by reviewing 50% of all epochs on consensus-hypnograms of the DOD database of 81 subjects (56 OSA + 25 healthy), where each PSG was scored by multiple experts. In our case, the level of this performance was achieved on ID as well as on two OOD single-scorer datasets of a considerably larger size containing subjects from a full spectrum of sleep-disorders. We consider results on our OOD datasets to be remarkably positive since the adaptation of the approach to the scoring taste of a single scorer is expected to be more difficult for algorithms (U-Sleep, confidence network) trained on data containing scorings of different physicians, as it represents a change of domain from multiple- to single-scorer ones. Adapting to the single-scorer's taste is closer to the current setup in clinical practice, where obtaining multiple-scorers' consensus is costly, and a single physician evaluates the PSG and makes the final clinical decisions. These results spotlight not only the efficacy of our approach and its robustness to OOD data with different diagnosis statuses but also underscore the potential to reduce the physicians' workload on manual sleep staging, which is paramount in practical scenarios.

Yet, our work is not without limitations. The field of uncertainty quantification for sleep staging is relatively new, and it does not include well-established baselines that would also incorporate publicly available data covering the full spectrum of sleep disorders. The data in the BSDB are mostly observational, i.e., subjects undergo sleep studies due to suspicion or symptoms, and so, the presence of different diagnoses is not randomized or balanced. The training of both classification and uncertainty-estimation algorithms was done without explicit control for gender, ethnicity, age, and clinical diagnosis, which may – together with non-randomized data – contribute to computational bias.

2.5 Conclusion

The significant challenges in automatic sleep staging, such as noise-amounts due to inter-scorer disagreement, and heterogeneity in PSG databases – reflecting the large inter-individual variability in sleep manifestation – underscore the complexities in achieving an AI model that could perfectly generalize to data from different domains. While automated sleep scoring algorithms have achieved excellent performances despite these hurdles, they are still bound by the limitations inherent to the quality of their training labels. Consequently, despite the technological advancements, the critical role of physicians in reviewing and verifying predicted hypnograms remains – so far – irreplaceable and imperative. With the increasing prevalence of sleep-wake disorders, and with the massive amounts of data present in PSGs, it is therefore necessary to drive research efforts to optimize physician's review by directing them to potential areas of uncertainty, while ensuring an efficient examination compliant with clinical needs.

In this study, we developed a pipeline aimed at enhancing the use of automated sleep-scoring algorithms in clinical practice. By retraining of the U-Sleep algorithm on 19,578 PSGs coming from 13 open-access databases, we reached state-of-the-art performance ($F1_w \geq 80.5\%$ on all test data) and encoded the sleep-scoring expertise of a broad range of physicians. Utilizing the comprehensive BSDB database of 8,832 additional PSGs, we compared various approaches for uncertainty quantification, including a novel confidence network that we designed to work in tandem with U-Sleep. Compared to softmax-based measures, our confidence network demonstrated its superiority for identifying predictions discordant from physician's scoring (AUROC $\geq 82.5\%$ on all test data) and built a prerequisite for successful implementation of a system that efficiently incorporates physician's insights.

Our study makes a significant contribution to sleep science by demonstrating the potential of incorporating a semi-automated approach into clinical settings. This is achieved through a unique combination of the U-Sleep robustness, the precision of an added confidence network, and the richness of the BSDB database, enabling in-depth validations with respect to individuals' diagnoses and accommodating the scoring preferences of different physicians. The combined approach of our pipeline ensures that while insights from the automatic sleep-scoring tool are utilized, physicians can concentrate their efforts on reviewing segments of biosignals where potential disagreements or algorithmic errors may occur. This has a great potential to significantly reduce the workload in the analysis of sleep studies. Moreover, the design of our pipeline can be applied beyond the sleep-scoring framework, for any use case where expert verification of algorithmic predictions is needed.

We believe that the adoption of scoring algorithms for clinical practice does not consist in replacing the physician's expertise with an algorithm, but mainly in enabling the effective use of the algorithm's insights and their thorough validation.

Chapter 3

Framework for Algorithmic Bias Quantification and its Application to Automated Sleep Scoring

Abstract

The validation of predictive algorithms is gaining importance with the increasing use of AI. Traditional validation of software seeking, for example, clinical certification involves correlation or Bland-Altman analysis comparing differences between predicted and reference values. However, such approaches are subject to simplifying assumptions on the algorithmic errors: normality of their distribution, homogeneity of variance, and independence from external factors. Our study, motivated by the sleep medicine use-case, proposes an in-depth quantification of systematic algorithmic error (*bias*) using the flexible statistical tool GAMLSS. Our approach allows the estimation of the bias distribution, identification of bias-generating factors, and extrapolation of various quantities assessing prediction validity.

Keywords:
Bias Quantification, Explainable AI, Model Validation, Automated Sleep Scoring

3.1 Introduction

Scoring of polysomnographic (PSG) recordings into 5 sleep-wake stages: Wake (W), Rapid-Eye-Movement (REM), and 3 Non-REM (N1, N2, N3), is a routine of many physicians. This process requires following AASM guidelines [8], and spending up to 2 hours determining the stage for every 30-second window (epoch) of a single-night data. Due to different and often subjective interpretations of the AASM guidelines, physicians reach inter-rater agreement of about 80% [15]. Research in automatic sleep scoring, motivated by high medical costs, has advanced since the 1960s [132]. In the last decade, Deep-Learning-based algorithms became prominent thanks to their ability to handle large datasets and capture complex patterns [48]. Yet, the algorithms trained on a broader range of PSG databases are in their performance technically limited by inter-scorer agreement. Hence, the classification accuracy of about 80% can be considered as near-perfect. An example of a state-of-the-art sleep-scoring algorithm is the U-Sleep [59], evaluated on the most extensive set of 16 clinical databases and reaching robust human levels of performance.

Automated tools have a great potential to enhance insights and improve physicians' efficiency. However, their transition to clinical use requires extensive validation - not only in classification performance but also in clinical relevance. Typically, algorithm performances are presented in terms of their average epoch-wise agreement metrics like the accuracy or the F1-score. While adequate for computer science, much more is needed for their clinical adoption. A subset of research validates algorithms also on a more clinically relevant subject-specific basis and evaluates how their performance correlates with individuals' demographics (e.g., age) or clinical status (e.g., rates of apneic/movement-events, AHI/PLMI) [133]. The more detailed validity assessments include Bland-Altman (BA) plots [134], comparing errors against reference values, typically quantifying their mean, and the magnitude

of their variability as a $\pm 1.96\times$standard-deviation (SD). Assuming a normal distribution of errors, such an approach intends to cover 95% of the expected error range. The BA-plot-derived validation typically involves clinically informative parameters (biomarkers), like AHI for apnea detectors or sleep stage percentages for sleep-scoring tools. Therefore, such validation is common in studies seeking clinical certification (e.g., [16], [135]).

Whether correlation or BA analysis is applied, it is crucial to highlight limitations (cf. [136]) that could cause the validation results to seem excessively positive to their intended user, the physician:

- *Correlation does not guarantee validity*; even systematically shifted - *biased* - predictions can be perfectly correlated with their reference.
- *BA plots assume errors to follow a normal distribution with homogeneous variance*, which may not always be true as variance often increases with the magnitude.
- *Both correlation and BA plots overlook the potential impact of any external factors.*

In this preliminary work, we demonstrate our framework while testing the cutting-edge sleep-scoring algorithm U-Sleep on a comprehensive clinical out-of-domain database. Specifically, we focus on using GAMLSS to identify potentially nonlinear biases, quantify their distribution, and examine their relation to relevant domain-specific factors. All this considering the ability of U-Sleep to predict one of the key prediction-derived markers of automatic sleep-scoring: the W%-state.

3.2 Materials and Methods

3.2.1 Bias and its quantification using GAMLSS

The term *bias* can generally be understood as a systematic deviation in estimates/predictions, $\hat{y}$, to their reference/true values, y. In statistical terminology, an estimate is called biased if the expected value of its error is non-zero [137]:

$$E(\hat{y} - y) \neq 0. \tag{3.1}$$

Typically, it refers to an estimation of parameters, $\hat{\theta}$, of a statistical model, $p(y|\theta)$. Transferably, bias is understood in a broader sense: *gender-bias* is said to be present if the output systematically differs in dependence on the perception of male/female or other genders, *age-bias* if the output is affected by the consideration of the individual's age, etc. Let's denote such factor(s) inducing biases as a variable x.

Combining statistical terminology and the comprehension of different sources of bias in a broader sense, we obtain:

$$E(\hat{y} - y) \sim x, \tag{3.2}$$

denoting the relation ($\sim$) between the bias (expected error) and the bias-inducing factors x. If $E(\hat{y} - y)$ is independent of x, or of their function $f(x)$, meaning the errors in predictions occur randomly, there is no x-induced bias. This framework already generalizes the validation assessments based on correlation analysis, $cor(\hat{y}, y)$, or more advanced BA plots, $(\hat{y} - y) \sim \frac{y+\hat{y}}{2}$, by considering external factors x and their relation to the error.

Considering more bias-potential factors, $(x_1, ..., x_k)$, being in an arbitrary functional form f to the expected error, the framework can be extended to:

$$E(\hat{y} - y) \sim f(x_1, ..., x_k). \tag{3.3}$$

As the f is unknown, one needs to make some assumptions to enable estimation of this relation: *bias-quantification (BQ)*. For example:

$$E(\hat{y} - y) \sim \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \epsilon, \tag{3.4}$$

yields linear-regression-based BQ, with standard assumptions such as $\epsilon \sim \mathcal{N}(0, \sigma^2)$. A significant β_0 can be interpreted as a baseline bias, similar to correlation- and BA-plot-based

validations, and any significant β_k as x_k-induced bias. This approach quantifies bias in a linear way (i.e., the effect of β_k is supposed to be the same regardless of the value of x_k), and assumes a homogeneous variance σ^2 independent of x. Importantly, it focuses on the estimation of the expected (mean) bias conditioned on x, rather than on its entire distribution.

The GAMLSS is a flexible statistical framework to model a broad range of probability distributions $\mathcal{D}$ using flexible models unique for their location (μ), scale (σ), and shape (λ) distributional parameters [138]. For the BQ specifically, the GAMLSS can be formulated as:

$$\hat{y} - y \sim \mathcal{D}(\mu = f(\boldsymbol{X}|\boldsymbol{\beta}), \sigma = g(\boldsymbol{X}|\boldsymbol{\gamma}), \lambda = h(\boldsymbol{X}|\boldsymbol{\zeta})),$$

where f, g, h stand for arbitrary functions of μ, σ, λ-specific predictors parameterized by $\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\zeta}$, respectively. For example, these functions may be an identity with linear predictors for linear regression, a link function for Generalized Linear Models (GLM), splines for Generalized Additive Models (GAM), but also a neural network. The choice of a distribution $\mathcal{D}$, depends on the support of an error, but we found that the BQ using an extended Normal distribution with flexible predictor functions for both μ and σ is a reasonable choice. The benefits of using GAMLSS for BQ are numerous, as it enables:

- *Quantification of the entire distribution of the bias*, i.e., not only of the mean but also of arbitrary quantiles,
- *Contribution assessment of external bias-inducing factors* x to individual distributional parameters using standard statistical tests (AIC, ANOVA, t-test, etc.),
- *Quantification of non-linear biases*, e.g., non-monotonic contribution of x (like age) to the bias using splines as predictor functions.

Based on the detailed knowledge of the bias distribution, one can extrapolate and fix x at arbitrary values of interest to *(i) quantify the expected value or arbitrary quantiles*. Based on that, one can also estimate *(ii) the expected probability that the bias lies within the expert-defined Region Of Practical Equivalence (ROPE)*, or *(iii) the probability that predictions are systematically over/under-estimating their reference*.

Despite the flexibility of GAMLSS, it is always up to the researcher to cope with the standard challenges of model building: identifying a suitable bias distribution, selecting relevant x, and choosing an appropriate functional form of predictors. Naturally, this choice is dependent on the application domain and the amount of available data, taking into account the balance between the complexity and practicality of the approach. Considering BQ as a framework for model validation and explainability, simpler approaches should be preferred.

3.2.2 Dataset

For the evaluations, we utilized the Berner Sleep Data Base (BSDB) from our partner clinic, Inselspital, University Hospital Bern [125]. From 2000 to 2021, more than 8000 PSG recordings were collected in individuals covering the entire spectrum of age (0-91 years) and sleep diagnoses. Signals were recorded at 200 Hz, and over the 20 years of data collection, manually scored according to AASM rules by one of more than the total of 60 physicians involved. For our work, we identified a subset of 4,075 PSGs with complete sleep scoring and available information on individuals' age, gender, AHI, and PLMI. This subset was used to characterize their demographic and clinical profiles and identify potential bias-inducing variables x

3.2.3 U-Sleep: the sleep scoring algorithm

The U-Sleep, introduced by Perslev et al. (2021) [59], is a deep convolutional neural network designed for sleep stage classification. It processes EEG-EOG channel pairs sampled at 128 Hz and outputs an array of softmax values, representing probabilities of the 5 sleep stages, for each signal window of the desired length. If more channel pairs are available, the U-Sleep implements *majority voting* and averages the predictions. The architecture includes an encoder-decoder part supplemented with skip connections and a classifier layer. The U-Sleep demonstrated state-of-the-art performance on 16 databases with over 15,000 participants, achieving an average F1-score of 79%. To date, this algorithm is cutting-edge

in terms of its performance as well as the amount and diversity of its training data. Furthermore, additional tests claim high resilience of U-Sleep and generalizability over different age groups of subjects [60]. For our evaluations on BSDB, representing out-of-domain test data, we used the most recent implementation of U-Sleep trained on 13 open-access databases of 19,578 PSGs, described as an experiment (i) on p. 3 in [60], reaching $(76.5 \pm 10.6)\%$ in F1-score on test data.

3.3 Results

3.3.1 Baseline performance of U-Sleep algorithm

The dataset of 4,075 subjects consisted of 1,504 females (F) and 2571 males (M) aged 0-86 and 0-91, respectively. The AHI/PLMI ranged from 0 to (141.6/131.5) for F and from 0 to (155.2/240.0) for M. The U-Sleep reached the mean $\pm$ SD in on-subject averaged-F1-score of $(74.29 \pm 11.07)\%$, and the mean W%-error of $(-1.13 \pm 5.15)\%$ and $(-1.45 \pm 6.05)\%$ for F and M, respectively. The natural question is whether such a difference in W% is systematic, only sex-related, or whether it can be attributed to particular distributions of age, AHI, and PLMI. Using the W% as a parameter of interest, we will demonstrate the application of our GAMLSS approach for BQ. The bias should be interpreted considering the expected W%, which according to the meta-analysis follows an increasing trend, from 2-3% for children to 20% for adults over 80 [17].

3.3.2 Bias Quantification (BQ)

Potential Bias-Inducing Factors: As factors x that may contribute to the bias, we considered gender, age, AHI, and PLMI. This choice was domain-specific and driven by evidence (e.g., [78]) that they highly influence individuals' structure of sleep. We aimed to evaluate whether, and eventually how much, these factors contribute to the algorithm's on-subject W%-bias. Based on that, we aimed to extrapolate the expected bias distribution conditional on arbitrary values of x.

BQ Model Class: To quantify the W%-bias, we modelled the difference between U-Sleep and physician-derived W% using the extended Normal distribution $N(\mu, \sigma)$, with separate μ and σ predictors. The μ, σ stand for location and scale distributional parameters directly related to the expected value and the SD, and fully identify distributional quantiles.

BQ Model Structure: To quantify whether and how much an arbitrary factor x contributes to the bias, each distributional parameter $\boldsymbol{\theta} = (\theta^{(1)} = \mu, \theta^{(2)} = \sigma)$ was modelled as:

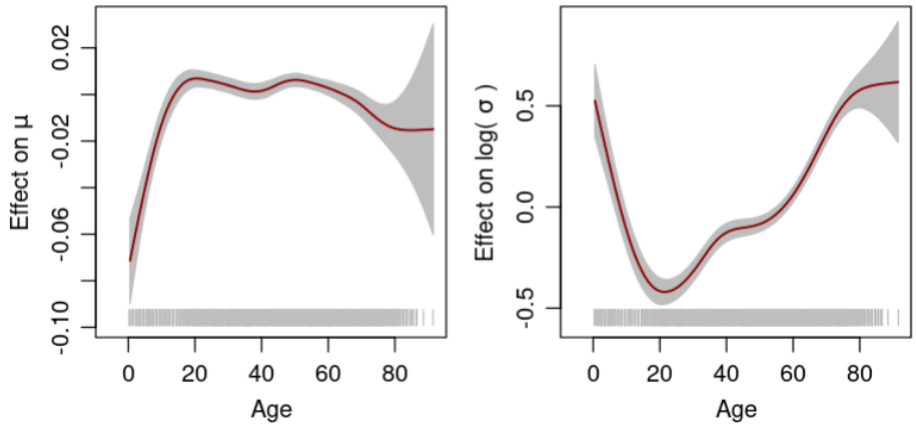
$$\theta^{(i)} = f(c_0^{(i)} + c_1^{(i)}\text{Gender} + c_2^{(i)}\text{AHI} + c_3^{(i)}\text{PLMI} + s(\text{Age}; c_4^{(i)})), \tag{3.5}$$

where $c_j^{(i)}$-s are $\theta^{(i)}$-specific coefficients and f is a suitable link function: the identity for μ and log for σ. We anticipated linear effects on gender, AHI, and PLMI (e.g., an increasing σ due to higher AHI), and a possibly nonlinear age-effect as quantified by the cubic spline s. To evaluate which factors contribute significantly to each parameter, a forward-building procedure was applied to Eq. 3.5, starting from the baseline predictor $\theta^{(i)} = f(c_0^{(i)})$, and iteratively adding the most significant term based on ANOVA (p-val < 0.05), if available.

Estimated W%-bias model: The following significant predictor functions were identified for the μ and σ parameters:

$$\begin{aligned} \mu &= -0.009 - 0.0001\ \text{AHI} - 0.0001\ \text{PLMI} + \text{s(Age)} \\ \log(\sigma) &= -3.845 + 0.058\ \text{GenderMale} + 0.005\ \text{AHI} \\ &\quad + 0.006\ \text{PLMI} + \text{s(Age)} \end{aligned} \tag{3.6}$$

Figure 3.1: Partial effects of age on the mean (μ) and standard deviation (σ) of the W%-bias model, quantified with cubic splines.

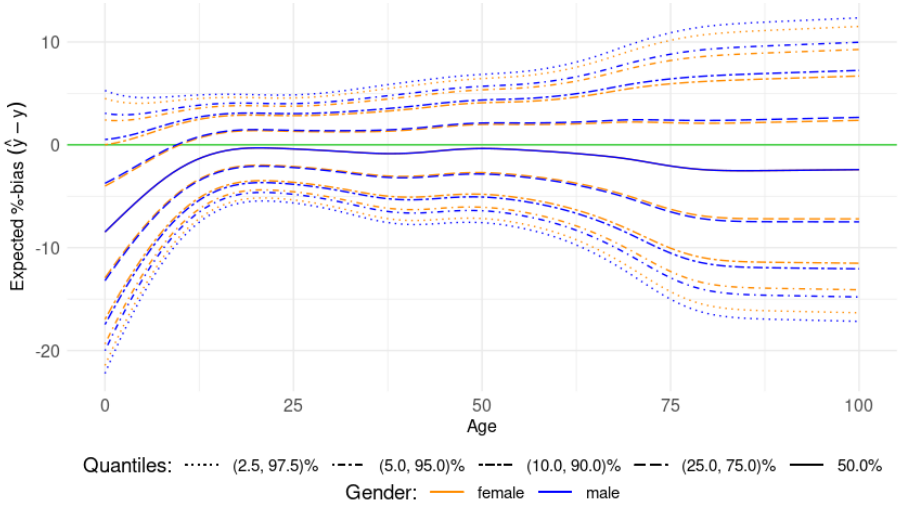


Interpretation: The baseline mean-bias (μ) of -0.009 refers to the U-Sleep underestimation of W% by 0.9%. This bias tends to expand with both AHI/PLMI and is not dependent on gender. In contrast, the baseline SD of the bias, $\sigma = 2.1\% = 100\% \times e^{-3.845}$, tends to increase for males and with AHI/PLMI. Both μ and σ are dependent on the non-linear effect of age identified by splines s, as depicted in Figure 3.1. The magnitude of the bias and its spread tend to be smaller for subjects between 20-60 years and to increase otherwise. Particularly, children under 10 tend to have greater underestimation of W% and an increased variability (e.g., an additional bias of about -0.07 for μ and $64.9\% = 100 \times (1 - e^{0.5})$ increase in σ for newborns). This important finding reflects the under-representation of children in the U-Sleep training data (only 1 of 13 databases, cf. [60]) as a likely source of this bias.

Derived Quantities: Quantifying bias distribution enables validation beyond the significance of the model's terms using a broad range of derived quantities: (i) bias quantiles, (ii) ROPE coverage, and (iii) probability of an over/under-estimation, as depicted in Figures 3.2,3.3, and 3.4, respectively. Based on Eq. (3.6), all figures extrapolate an optimistic ('*healthy*') scenario of AHI = PLMI = 0 and illustrate outcomes subject to an individual's age and sex. To better demonstrate our results, we created an interactive **app** showing the results also for other sleep stages and providing technical details on underlying bias-models.

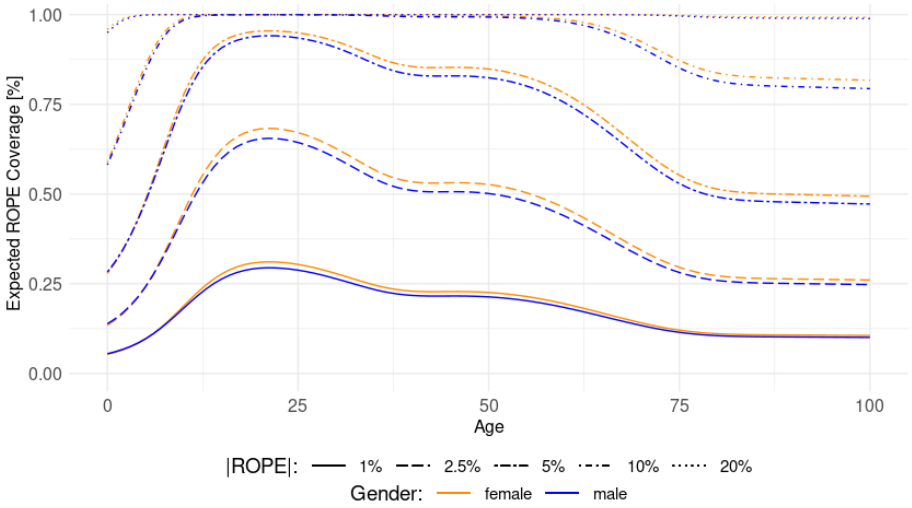
Figure 3.2 shows that for arbitrary age and gender, the W% is underestimated, as the median-bias < 0, with the greatest bias of -8% for newborns. Bias magnitude and its variability increase for younger and older subjects. Further, more than 25% (lower dashed lines) of subjects aged under 10 and above 70 have the W% underestimated by at least 5%.

Figure 3.2: Expected quantiles of the W%-bias.



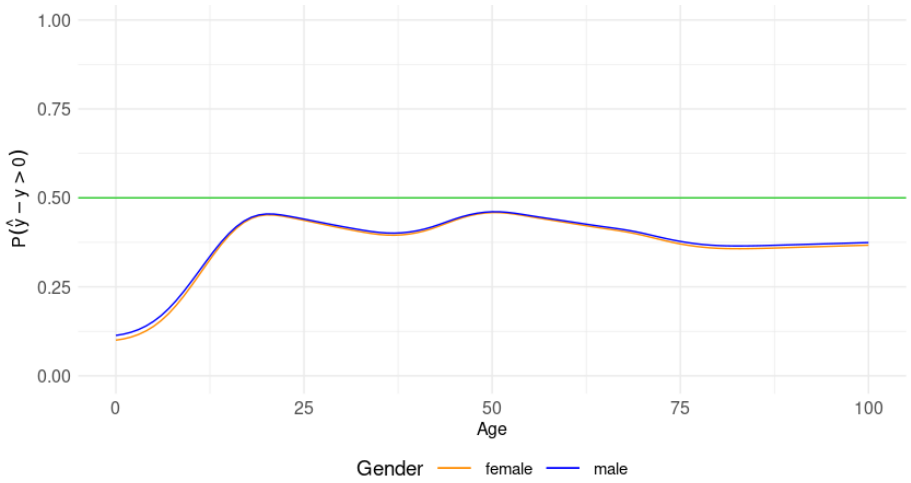
As depicted in Figure 3.3, considering the error-magnitude of 2.5% to be irrelevant and to define the ROPE, the algorithm predictions would cover about 15-20% of babies, 65% of adults aged 20, and up to 30% of adults above 75.

Figure 3.3: Expected coverage of the region of practical equivalence (ROPE) for W%-predictions, expressed as an interval of ±ROPE thresholds.



Finally, Figure 3.4 depicts the probability of over-estimated ($\hat{y} - y > 0$) predictions. The green horizontal line depicts a situation of a symmetric bias distribution where the median would be 0, and the chances of over- and under-estimation equal. In the case of W%, the predictions are for both genders systematically underestimated: for >80% of babies and 55-60% of adults between 20-65 years.

Figure 3.4: Probability of positive bias (overestimation, $\hat{y} - y > 0$) in W%-predictions.



3.4 Discussion & Conclusion

Detailed validation of predictive algorithms is essential for in-depth evaluation of their accuracy, and it is a prerequisite for their successful adoption. Our study presents a universal framework for flexible quantification of the algorithmic bias, which may be in a possibly non-linear relation to external controlling factors. We illustrated this framework on the use-case of the cutting-edge sleep-scoring algorithm U-Sleep, focusing on W%ake estimation, which is essential for assessing sleep efficiency. We outlined the steps of bias-model building, identification of relevant bias-contributing factors, and extrapolation of results through various quantile-based metrics. As an important result, we revealed a bias for children who were underrepresented in the original training data of the U-Sleep, illustrating the practicality of our approach for detecting gaps in the training data or their insufficient heterogeneity. Results of our study are also available within an **app**, allowing interactive visualizations of results and their discussion with domain experts.

We are convinced that with the growing use of complex AI models across various domains, research into approaches for their flexible and explainable validation should not be delayed.

Acknowledgment

The secondary usage of BSDB from Inselspital, University Hospital Bern, was ethically approved (KER-Nr. 2022-00415).

Chapter 4

Beyond Accuracy: A Framework for Evaluating Algorithmic Bias and Performance, Applied to Automated Sleep Scoring

Abstract

Recent advancements in artificial intelligence (AI) have significantly improved sleep-scoring algorithms, bringing their performance close to the theoretical limit of approximately 80%, which aligns with inter-scorer agreement levels. While this suggests the problem is technically solved, clinical adoption remains challenging due to ethical and regulatory requirements for rigorous validation, fairness, and human oversight. Existing validation methods, such as Bland-Altman analysis, often rely on simple correlation metrics, overlooking potential non-linear influences of external factors (e.g., demographic or clinical variables) on systematic predictive errors (biases) in derived clinical markers. Additionally, performance metrics are typically reported as the mean of on-subject results, neglecting critical scenarios—such as different quantiles—that could better convey the algorithm's capabilities and limitations to clinicians as end-users. To address this gap, we propose a universal framework for quantifying both performance metrics and biases in predictive algorithmic tools. Our approach extends conventional validation methods by analyzing how external factors shape the entire distribution of predictive performance and errors, rather than just the expected mean. Applying it to the widely recognized U-Sleep and YASA sleep-scoring algorithms, we identify biases—such as age-related shifts—indicating missing input information or imbalances in training data. Despite these biases, we illustrate that both algorithms maintain non-inferior performance in the risk assessment of sleep apnea based on prediction-derived markers, highlighting the potential and clinical utility of algorithmic insights.

4.1 Introduction

Polysomnography (PSG) is the gold standard for diagnosing sleep disorders, offering comprehensive insights into sleep architecture through multi-channel recordings, including brain activity (electroencephalography, EEG), eye movements (electrooculography, EOG), and muscle tone (electromyography, EMG). Scoring each 30-second window of PSG recordings into five discrete sleep-wake stages—wake (W), non-rapid eye movement stages 1-3 (N1, N2, N3), and rapid eye movement (REM) sleep—is traditionally done by expert physicians following the American Academy of Sleep Medicine (AASM) guidelines [8]. Such structured scoring, known as a *hypnogram*, allows for the extraction of various sleep markers, such as REM-latency and sleep efficiency, which are crucial for evaluating sleep quality and identifying potential sleep disorders [18], [139].

However, manual sleep-scoring is time-consuming, often requiring hours to evaluate a single night's data, and is subject to variability between scorers. Notably, the inter-rater agreement among human scorers reaches, depending on metric used, approximately 75-80%[14], [15], [66], [140], setting a technical upper bound on the performance of automated

algorithms trained on datasets containing reference labels from multiple scorers [80], [81], [141]. Despite these limitations, sleep-scoring algorithms have shown a great potential for reducing the manual workload of clinicians, with widely recognized tools such as the deep learning-based U-Sleep [59], [60] and the machine learning-based YASA [58]. U-Sleep, utilizing a deep convolutional neural network, reported a macro F1-score of 79% [59]. In contrast, YASA employs a lightweight approach, extracting time- and frequency-domain sleep-related statistical features from PSG signals, making it computationally less demanding. On a validation set of 25 healthy adults scored by five independent scorers, YASA reported a mean (interquartile-range) macro F1-score of 78.5% (9.4), compared to 82.7% (7.7) for U-Sleep [58]. In a cohort of 50 adults with sleep apnea, YASA scored 70.1% (15.5), while U-Sleep reached 78.7% (10.9) [58]. Both validation cohorts of 25 + 50 subjects reported in [58] were sourced from the multi-scorer DOD database [70]. Despite methodological differences, both algorithms perform close to inter-rater agreement levels.

While artificial intelligence (AI) holds great promise for automating sleep-scoring, the inherent variability in human labels used to train these tools imposes a theoretical performance ceiling [141]. Due to about 20-25% noise level in sleep-stage labels inherent from inter-scorer disagreement [14], [15], [66], [140], the algorithms trained on a broad range of databases containing scoring tastes of multiple human-annotators technically struggle to exceed the performance-level of about 75-80%, aligning with inter-rater agreement [80], [81], [142]. In consequence, even the state-of-the-art systems will inevitably show some level of error relative to individual scorers, and a single physician is expected to disagree with approximately 20-25% of AI predictions, reflecting the fundamental challenge in automated sleep scoring and its integration into clinical practice [95]. This is a critical limitation, as physicians remain responsible for both the scoring and diagnostic decisions that follow, raising ethical and practical concerns. In response, emerging legal frameworks such as the EU AI Act (Regulation (EU) 2024/1689) and the Medical Device Regulation (MDR, Regulation (EU) 2017/745) have been introduced to ensure transparency, fairness, and human oversight in AI-driven (healthcare) solutions. While the U.S. Food and Drug Administration (FDA) has not yet established a dedicated AI regulatory framework, it has issued guidance on AI/ML-enabled medical devices, including the *Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices* resource and the *Artificial Intelligence-Enabled Device Software Functions* draft guidance, which align with the principles of the EU AI Act and MDR.

Despite significant improvements in algorithmic performance, the clinical adoption of AI-based sleep-scoring tools—such as U-Sleep and YASA—remains limited. This may be due to a combination of factors, including regulatory challenges [85], such as the requirement for human oversight (cf. EU AI Act), and the current lack of deployed clinical pipelines to support such integration efficiently [95]. Additional barriers include limited clinician trust and lack of interpretability [86], [87], possible ambiguities in the interpretation of AASM scoring guidelines [11], [15], [88], and signal artefacts that contribute to substantial inter-scorer variability [8], [15], [66], [88], [140]. Concerns about *algorithmic bias* [91], [96], where predictions systematically vary across demographic or clinical subgroups, and broader issues of fairness and equity in clinical AI applications [89], [90] also play an important role.

Beyond legal and ethical considerations, current validation methods—typically relying on average (possibly on-subject) accuracy or correlation-based metrics such as Bland-Altman (BA) plots [134]—often fail to assess clinical reliability fully and may even lead to misinterpretation [136]. These methods overlook important characteristics of expected model performance and error distribution (e.g., min, max, and different quantiles), assume normality and uniform variance across the entire testing population, and typically disregard the influence of external factors (e.g., the dependence of algorithmic errors on demographics). Additionally, correlations can obscure systematic biases, as even systematically shifted - *biased* - predictions can be perfectly correlated with their reference values. Therefore, there is a growing need for more comprehensive validation approaches that evaluate whether AI models perform consistently across diverse (sub)populations and offer clinically meaningful, reliable, and granular insights into the distributional characteristics of algorithmic performance and errors [96].

In this study, we extend our recently proposed framework for quantifying algorithmic bias in predictive AI tools [96], with a specific focus on automated sleep-scoring. We systematically evaluate the performance of two widely recognized models, U-Sleep and YASA, and

assess the validity of clinically relevant markers (e.g., REM latency, sleep efficiency) derived from their predicted hypnograms. These markers represent key metrics routinely included in PSG reports to support diagnostic decisions. Leveraging a large, out-of-domain clinical dataset, we examine whether these algorithms exhibit biases across demographic and clinical factors, quantifying both the magnitude and distributional characteristics of their performance metrics and errors in predicting derived PSG markers. Key contributions of this study include:

- *A general framework for algorithmic bias quantification* applied to the context of automated sleep-scoring,
- *Systematic evaluation of algorithmic performance*, including expected distributions and systematic shifts across demographic and clinical subgroups,
- *Quantification of algorithmic biases* in prediction-derived clinical PSG markers, including bias distributions and their dependence on external factors,
- *Assessment of the diagnostic utility of potentially biased markers*, particularly in the detection of obstructive sleep apnea,
- Provision of an interactive R-Shiny app for dynamic exploration of results.

By proposing a framework for in-depth *bias quantification*, we aim to ensure that AI-based clinical tools are not only accurate, but also fair, reliable, and transparently validated. Understanding algorithmic performance and its limitations is essential for bridging the gap between computer scientists developing these tools and clinicians as end users [143]. This collaboration can be strengthened by increasing awareness of AI's potential benefits while managing expectations through clear reporting of its capabilities and limitations. While we demonstrate our framework on U-Sleep and YASA, this study does not aim to compare which algorithm performs better, as they were trained on different datasets. Rather, our primary objective is to present the framework as a tool for communicating algorithmic performance and error distributions to end users (clinicians) and for use within clinical certification processes to evaluate model reliability across diverse clinical contexts.

4.2 Materials and Methods

In this section, we review common methods for assessing predictive performance and present our framework for systematically quantifying algorithmic bias and performance [96]. Finally, we introduce a study use-case demonstrating this framework by evaluating the sleep-scoring algorithms U-Sleep and YASA on a large, out-of-domain clinical dataset.

4.2.1 Current standards and limitations in reporting algorithmic performance

Predictive AI (statistical, machine- or deep-learning) models are subject to multiple sources of uncertainty [144], including model architecture (e.g., predictor selection in regression; number of hidden layers in neural networks, NN), label noise (e.g., inter-scorer disagreement in sleep stage annotations), and dataset heterogeneity (e.g., demographic imbalances or variability in recording devices). Additionally, NNs are prone to overconfidence [127]—assigning excessively high probabilities to their predictions—and often rely on spurious correlations rather than causal relationships, leading to systematic biases [145].

Therefore, despite the fast and often accurate insights predictive tools can provide, it is essential to rigorously validate them [146]. The classification performance is commonly reported using aggregated metrics such as accuracy or F1-score, typically averaged across all epochs (windows/observations) of the test data. Particularly for healthcare, a more informative approach is to report subject-wise (test) performance, supplemented with variability measures such as standard deviation (SD) to capture inter-subject differences. For numerical outcomes in a regression setting (e.g., predicting the Apnea-Hypopnea-Index, AHI, capturing frequency of breathing arrests in apnea-detectors), validation is often based on the correlation between predicted ($\hat{y}$) and reference (y) values, even though the concepts of correlation

and validity are quite distinct [147]. Particularly for tools seeking clinical certification, this is extended with Bland-Altman (BA) plots [134], which compare the prediction errors ($\hat{y} - y$) against reference values and typically approximate the error distribution as the mean error rate ± a multiple of the SD.

However, these conventional validation methods have critical limitations [136], [146], [147]. Classification metrics alone typically fail to report performance variability across subpopulations (e.g., for different ages, genders, or diagnoses) and do not evaluate critical scenarios including minimum/maximum and different quantiles in algorithmic performance or error rates, also concerning subjects' characteristics. Moreover, regression tools that are validated using correlation or BA analysis implicitly assume symmetric, homoscedastic errors independent of external factors, which is a rare case in practice. Notably, correlation-based assessments may often fail to detect systematic biases, as the correlation can be perfect: $cor(\hat{y}, y) = 1$, even if predictions are systematically shifted: $\hat{y} = y + c$. These limitations highlight the need for a more comprehensive validation framework that extends beyond clinical AI applications, offering a detailed assessment of algorithmic performance, error distribution, and their dependence on external factors.

4.2.2 Framework for algorithmic bias quantification using GAMLSS

Bias, in the context of model validation, refers to the systematic deviation between model predictions ($\hat{y}$) and true reference values (y). Conventionally, a prediction (or estimate) is considered *biased* if the expected value of its error deviates from zero, i.e., $E(\hat{y} - y) \neq 0$ [137]. In practice, this *systematic deviation* (= *bias*) may stem from multiple factors such as age, gender, or other attributes, which we collectively denote as *bias-inducing variables* (X) or sensitive attributes [91]. These factors influence predictive errors and performance, leading to systematic shifts in their distribution that can be modeled and quantified.

In previous work [96], we introduced a framework for *Bias Quantification* (BQ), which extends beyond measuring baseline bias, $E(\hat{y} - y)$, to assess the influence of external factors, such as age or gender. Specifically, we proposed quantifying the relationship between predictive errors and bias-inducing variables X through conditional expectation:

$$E(\text{bias}) = E(\hat{y} - y) \sim f(x_1, \ldots, x_k|\theta), \tag{4.1}$$

where f represents an arbitrary function parameterized by θ, capturing the relationship between bias and the external factors $X = (x_1, \ldots, x_k)$. This framework extends traditional validation approaches, such as correlation- or Bland-Altman (BA) analysis, by systematically integrating external variables, providing a more comprehensive understanding of bias.

To operationalize this framework, we utilize *Generalized Additive Models for Location, Scale, and Shape* (GAMLSS) [138], a highly flexible statistical approach that models the entire distribution (of bias) rather than just its mean (as done in standard regression setting). The GAMLSS allows capturing not only the expected bias (location), but also its variability (scale), and other shape-related distributional properties, such as skewness, kurtosis, or inflation [138]. Each of these distributional parameters can depend on X in possibly non-linear ways. Specifically, we suggest to exploit GAMLSS to model the distribution of the bias as:

$$\text{bias} \sim \mathcal{D}(\mu = f_\mu(X|\beta_\mu), \sigma = f_\sigma(X|\beta_\sigma), \nu = f_\nu(X|\beta_\nu), \tau = f_\tau(X|\beta_\tau)), \tag{4.2}$$

where μ, σ, and ν, τ represent the location, scale, and two shape parameters of a chosen distribution $\mathcal{D}$ [148]. The predictor functions f_θ define the relationships between these parameters and the external factors X, which can be specified as linear, non-linear, spline, or even neural network-based, depending on the complexity of dependencies. The predictor functions for each $\theta \in \{\mu, \sigma, \nu, \tau\}$ are parameterized by $\boldsymbol{\beta_\theta}$, respectively.

Bias Quantification (BQ): As a suitable distributional choice for BQ, evaluating the distribution of algorithmic errors, we suggest using the generalized normal distribution:

$$\hat{y} - y \sim \mathcal{N}(\mu = f_\mu(X|\beta_\mu), \sigma = f_\sigma(X|\beta_\sigma)) \tag{4.3}$$

with separate predictors for its location (μ) and scale (σ) parameters. This BQ model effectively addresses the limitations of standard validation approaches by capturing both systematic biases (through μ) as well as their variability (i.e., heteroscedasticity, through σ) across different subgroups defined by bias-inducing factors X. The possibly non-linear impact of X on both μ and σ may be quantified through flexible functions (e.g., splines), enabling the estimation of arbitrary error quantiles for a more detailed bias characterization.

Performance Quantification: Beyond bias assessment, our framework extends to the modelling of performance metrics (e.g., accuracy and F1-score), accounting for their trends and variability across subgroups, defined by X. Since performance metrics are typically bounded between 0 and 1 (i.e., 0-100%), we propose using the zero-and-ones-inflated beta distribution ($\mathcal{B}$) [148]:

$$\text{Performance Metric} \sim \mathcal{B}(\mu = f_\mu(X|\beta_\mu), \sigma = f_\sigma(X|\beta_\sigma), \nu = f_\nu(X|\beta_\nu), \tau = f_\tau(X|\beta_\tau)), \quad (4.4)$$

which is well-suited for bounded data with potential boundary inflation. This distribution is characterized by four parameters: location (μ), scale (σ), zero-inflation (ν), and one-inflation (τ). By leveraging $\mathcal{B}$ and flexible predictor functions, our framework captures both systematic patterns in central tendency, and variability, and addresses extreme performance outcomes (such as complete misclassification or perfect classification), concerning arbitrary demographic or clinical characteristics X.

Key Advantages: Our framework offers several key advantages that are shared between bias and performance quantification:

- *Modelling of the full distribution*: Rather than focusing solely on average bias or performance, our approach models the entire distribution, allowing the estimation of arbitrary quantiles. This enables a comprehensive view of both bias and performance metrics across arbitrary subject characteristics (sensitive attributes) X included.
- *Capturing non-linear relationships*: By incorporating flexible, non-linear predictors (e.g., splines), the framework effectively captures complex, non-monotonic relationships between X and the bias or performance metrics.
- *Hypothesis testing*: The framework supports standard statistical hypothesis testing (e.g., using AIC, ANOVA, or t-tests) to assess the contribution of specific characteristics X.

4.2.3 Study use-case

Our framework is demonstrated to two recognized sleep-scoring algorithms: U-Sleep and YASA. We evaluate the validity of their predictions on a large, out-of-domain clinical dataset and investigate how demographic and clinical factors influence the distribution of performance metrics and the bias concerning prediction-derived clinical markers. In addition, we evaluate the diagnostic utility of (possibly biased) predicted markers in distinguishing between healthy subjects and OSA patients. The OSA, estimated to impact up to 17% of the general population [149], is the most prevalent sleep disorder and a significant risk factor for the development of cardiac events and overall mortality [150].

Sleep-scoring classifiers: U-Sleep and YASA

U-Sleep: Developed by Perslev et al. (2021) [59], U-Sleep is a deep learning-based model utilizing a convolutional neural network for multi-channel sleep staging. It processes EEG-EOG pairs, sampled at 128 Hz, and predicts the probability of five sleep stages (Wake, N1, N2, N3, REM) for each signal window of the specified length. When multiple channel pairs are available, U-Sleep aggregates predictions using *majority voting* that averages probability outcomes of individual pairs. Its architecture combines an encoder-decoder network with skip connections, followed by a classification layer. Originally evaluated across 16 large-scale sleep datasets (∼15,000 participants), U-Sleep reported the weighted-mean F1-score of 79% across evaluated datasets [59]. In our study, we use the latest implementation, trained

on 19,578 PSGs from 13 public datasets, described as an experiment (i) on p. 3 of Fiorillo et al. (2023) [60], which corrected a prior channel-derivation bug and reported a mean (standard deviation) F1-score of 76.5%(10.6) on unseen test data. U-Sleep has demonstrated strong generalizability across different age groups and sleep disorder profiles [59], [60].

YASA: Developed by Vallat and Walker (2021) [58], YASA employs a machine learning-based approach, utilizing LightGBM (Light Gradient Boosting Machine) for sleep stage classification. For each channel and 30-second window, YASA extracts statistical and spectral features (e.g., kurtosis, skewness, power ratios) from EEG, EOG, and optionally EMG signals. It was trained on 31,000 hours of data from 3,163 full-night PSGs across seven diverse datasets, covering a broad age range of mean (SD) of 49.8 (26.4) years and various sleep disorders. On a hold-out test set of 25 healthy subjects, YASA achieved a mean (inter-quartile-range) macro F1-score of 78.5% (9.4), and on 50 patients with obstructive sleep apnea (OSA), it achieved 74% (10.8), both sourced from the Dreem Open Datasets (DOD) [70]. While slightly outperformed by the deep learning-based U-Sleep—macro F1-score of 82.7% (7.7) and 78.7% (10.9) on DOD healthy and OSA, respectively [58]—YASA's lightweight architecture enables efficient large-scale processing in under five seconds per night. To ensure a fair comparison with the U-Sleep majority-voting mechanism, we averaged YASA's predictions across all EEG-EOG pairs of each PSG.

Data

We conducted our evaluations using the *Berner Sleep-Wake Registry* (BSWR), provided by Inselspital, University Hospital Bern. The database contains over 8,000 PSG recordings collected between 2000 and 2021 from predominantly symptomatic subjects (< 1% are healthy controls), mostly males (63.1%), across a broad range of age (0–91 years) and sleep disorders (e.g., OSA, narcolepsy). PSG signals were recorded at 200 Hz and manually scored by one of over 60 physicians according to the AASM guidelines [8]. To align older Rechtschaffen and Kales [75] scorings with the AASM standard, N3 and N4 stages were merged.

For our analysis, we selected a subset of 4,075 PSGs with complete sleep-scoring and available demographic and clinical data, including age, gender, Apnea-Hypopnea Index (AHI), and Periodic Limb Movement Index (PLMI). AHI was computed using the AASM "recommended" definition, requiring $\geq$30% airflow reduction with $\geq$3% desaturation or arousal for hypopneas. PLMI excluded limb movements associated with respiratory events or arousals. These were considered potential bias-inducing factors X (cf. Eq. 4.1) due to their known effects on sleep architecture and association with the majority of sleep comorbidities [18], [139], [151], [152]. Specifically, age and gender were included as demographic variables known to influence sleep physiology, while AHI and PLMI were used to characterize clinical subgroups, as both relate to breathing and movement disturbances that may introduce signal artefacts and can also alter sleep architecture [18], [152].

To quantify bias and performance, we utilized three sets of hypnograms: a "true" reference scored by physicians and two algorithmically predicted hypnograms from YASA and U-Sleep, all at a 30-second resolution. Using these hypnograms, we computed subject-specific accuracies, macro-F1 scores, and clinically relevant markers (e.g., sleep-stage percentages, sleep efficiency). The resulting dataset was structured into a tabular format with 4,075 rows, each representing a single PSG recording, while columns contained performance metrics, clinical markers, and bias-inducing variables X (age, gender, AHI, PLMI). This structure enabled a systematic analysis of patterns in both performance metrics and biases in hypnogram-derived clinical markers. Lastly, to evaluate the diagnostic utility of potentially biased markers, we included clinically conclusive diagnoses (e.g., OSA) as part of the dataset.

Ethics Approval and Consent The secondary usage of the Berner Sleep-Wake Registry (BSWR) dataset was approved by the local ethics committee (Kantonale Ethikkommission Bern [KEK]-Nr. 2022-00415), ensuring compliance with the Human Research Act (HRA) and Ordinance on Human Research with the Exception of Clinical Trials (HRO). All methods were carried out in accordance with relevant guidelines and regulations. Written informed

consent was obtained from all participants, as part of the general consent process introduced at Inselspital in 2015. Data were maintained with confidentiality throughout the study.

4.3 Results

4.3.1 Implementation

The analytical part of this study was conducted using the statistical software R and its development environment, RStudio [153]. To implement our framework and quantify the distribution of algorithmic performance metrics and biases concerning derived clinical PSG markers, we used the `gamlss v5.4-22` package. The evaluation of the diagnostic utility of derived markers was performed using `caret v6.0-94` and `glmnet v4.1-8`. Additionally, demographically balanced cross-validation splits were achieved using the anticlustering implemented in `anticlust v0.8.7`. To demonstrate our framework, we used the existing implementation of U-Sleep from the experiment (i) on p.3 of Fiorillo et al. (2023) [60] and YASA by Vallar and Walker (2021) [58] from the open-source python library `yasa v0.6.4`.

R Shiny App To supplement our findings, we provide an interactive web application built using `shiny v1.9.1`, allowing users to explore bias and performance quantification across demographics (age, gender) and clinical indices (AHI, PLMI), for both U-Sleep and YASA. The app is freely accessible at: https://mystatsapps.shinyapps.io/bias/, and features five main tabs: 1. `Expected Quantiles`: Displays the expected distributions of selected performance metrics (accuracy, F1-score) or bias for selected prediction-derived PSG marker, as a function of age, gender, and selected AHI/PLMI values. It also presents a table of expected quantiles at (1, 2.5, 5, 25, 50, 75, 95, 97.5, 99)% for both males and females. 2. `Region of Practical Equivalence (ROPE)`: Illustrates ROPE coverages for predefined sample thresholds to assess the expected proportion of predicted PSG markers within predefined ROPE-bounds of negligible errors (i.e., practical equivalence). 3. `Probability of Bias`: Computes the percentage of cases where the algorithmic prediction ($\hat{y}$) overestimates the physician-based reference value (y). 4. `Model Summary`: Provides detailed statistical outputs, including ANOVA tables, p-values, and other relevant information for each bias/performance model. 5. `Partial Effects`: Visualizes the effects of numeric bias-inducing variables (age, AHI, PLMI) along with their 95% confidence intervals. This is particularly relevant for age, whose effect was modelled as possibly a non-linear spline term. The app offers a flexible exploration of bias and performance quantification models, supporting further interpretability and reproducibility of our findings.

4.3.2 Descriptive statistics of bias-inducing variables

For both bias- and performance-quantification models (Eq. 4.3-4.4), we consider age, gender, AHI (Apnea-Hypopnea Index), and PLMI (Periodic Limb Movement Index) as potential bias-inducing variables X. These variables are clinically known to affect sleep and were therefore considered to eventually impact the predictive performance and the bias in sleep-scoring algorithms. The study dataset consists of a subset of 4,075 PSG recordings from Berner Sleep-Wake Registery (BSWR), where these variables were fully observed. Most recordings were from male subjects (63.1%) with a *mean (SD)* age of 50.1 (18.0) years, while female subjects accounted for 36.9% with a mean (SD) age of 45.9 (18.7) years.

The upper part of Table 4.1 presents further statistical characteristics of X, including Spearman correlations (ρ) and gender differences assessed by the Wilcoxon test. The overall mean age was 48.6 (18.4) years, with a broad range of 0-91 years. AHI, a measure of breathing arrest frequency, had a mean of 18.1 (20.1) with values ranging from 0 to 155.2. PLMI, reflecting the limb movement frequency, exhibited a mean of 13.4 (24.5), ranging from 0 to 240. Correlation analysis using Spearman's ρ, which robustly accounts for monotonic relationships, showed significant positive associations between age and both AHI ($\rho = 0.41$) and PLMI ($\rho = 0.27$). AHI and PLMI were also positively correlated ($\rho = 0.11$), supporting prior findings that breathing- and movement-related sleep disruptions tend to co-occur and increase with age [154]. Gender differences were evident in all numeric X variables.

Males—who were in our dataset older (median difference = 4)—exhibited a higher AHI (median difference = 7.3) and PLMI (median difference = 1.9), compared to females. These differences may be attributed either to the greater prevalence of sleep-disordered breathing in males [149] or to age-related increases in AHI and PLMI [154], given that males were older.

4.3.3 Algorithmic Performance

Descriptive statistics of performance metrics The bottom part of Table 4.1 summarizes the on-subject performance metrics, specifically accuracy and macro F1-score of U-Sleep and YASA achieved on BSWR, representing the out-of-domain dataset. Performance was evaluated using accuracy and macro-F1 score, calculated for the five-class sleep staging task, with the macro-F1 score computed as the unweighted average of the per-class F1-scores. U-Sleep achieved a mean (SD) accuracy of 79.2% (10.0), outperforming YASA, which reached 74.6% (11.8). The same trend was observed for the F1-score, with U-Sleep at 74.3% (11.1) versus YASA's 66.1% (13.3).

Performance varied substantially across subjects, with both models occasionally producing either perfect predictions (100%) or near-complete misclassifications (0%). These extremes justify modelling performance using a zero-and-ones-inflated beta distribution (Eq. 4.4) to capture both central and boundary tendencies. Looking at the lower quartile (Q25), U-Sleep maintained the accuracy and F1-score of 75.0% and 69.9%, while YASA dropped to 69.3% and 60.1%, indicating predictive power lower than the commonly reported inter-rater agreement level of about 75-80%[14], [15], [66], [140] for at least 25% of subjects, for both algorithms. Conversely, at the upper quartile (Q75) in accuracy and F1-score, the U-Sleep reached 86.0% and 81.7%, whereas the YASA 82.9% and 75.5%, respectively, both approaching or slightly exceeding the commonly reported inter-scorer agreement level in the literature [14], [15], [66], [140]. Given that our clinical data include scoring patterns from over 60 physicians, this suggests that both algorithms—based on naive sample-based comparisons ignoring influence of bias-inducing variables—achieve or exceed the performance bound of human-level agreement in about 25% of cases.

Further, correlations revealed negative associations between age, AHI, PLMI, and both accuracy and F1-score across models. For example, age was negatively correlated with U-Sleep and YASA's accuracies (ρ of -0.29 and -0.27, respectively), suggesting performance declines in older subjects. Similarly, AHI and PLMI negatively correlated with both performance metrics, indicating that individuals with more breath- and movement-related events (observed more in older) pose classification challenges, with steeper performance decline for AHI. In addition, the Wilcoxon test revealed gender differences, with both algorithms showing median performance reductions across metrics of about 2.5% in males.

These results highlight the importance of considering demographic and clinical factors in model evaluation. The substantial variability observed across different subgroups underscores the need for robust performance quantification frameworks, ensuring both, the transparent reporting of algorithmic capabilities, and their equitable deployment across diverse populations.

Table 4.1: Summary statistics of demographic and clinical variables and performance metrics for U-Sleep and YASA.

Metric	Mean	SD	Q10	Q25	Q50	Q75	Q90	Min	Max	ρ(Age)	ρ(AHI)	ρ(PLMI)	M - F
Demographic and Clinical Variables													
Age	48.6	18.4	23.0	36.0	51.0	62.0	72.0	0.0	91.0	1	0.41	0.27	4
AHI	18.1	20.1	1.4	4.1	11.0	24.4	44.8	0.0	155.2	0.41	1	0.11	7.3
PLMI	13.4	24.5	0.0	0.0	2.9	14.8	42.3	0.0	240.0	0.27	0.11	1	1.9
Model Performance Metrics													
Accuracy (U-Sleep)	79.2	10.0	66.8	75.0	81.4	86.0	89.2	1.0	100.0	-0.29	-0.35	-0.19	-2.18
Accuracy (YASA)	74.6	11.8	59.0	69.3	77.2	82.9	86.2	0.0	100.0	-0.27	-0.36	-0.19	-2.64
F1-score (U-Sleep)	74.3	11.1	59.9	69.9	76.9	81.7	85.2	1.7	100.0	-0.28	-0.27	-0.16	-2.23
F1-score (YASA)	66.1	13.3	47.7	60.1	69.4	75.5	79.6	0.0	100.0	-0.29	-0.29	-0.18	-2.69

Notes: Mean, standard deviation (SD), (10, 25, 50, 75, 90)%-quantiles (Q10, Q25, Q50, Q75, Q90), minimum (Min), and maximum (Max) values are reported. Spearman correlations (ρ) with age, AHI, and PLMI are provided, along with the median difference between males and females (M - F), assessed using the Wilcoxon test. Significant correlations and differences are highlighted based on the p-value thresholds: 0.05 , 0.01 , and 0.001 , respectively.

Performance-quantification models To assess the systematic impact of demographic and clinical factors on algorithmic performance (accuracy, F1-score), we employed the zero-and-ones-inflated Beta distribution (Eq. 4.4) within the GAMLSS framework. This approach enabled flexible modelling of performance distribution, rescaled to the [0,1] interval, capturing both expected quantiles and extreme cases of perfect (100%) or zero performance. We did not include Cohen's Kappa, a common metric of agreement comparison, in our analysis due to its value range of [–1, 1], which would complicate direct interpretability and comparison with other [0,1] bounded metrics (accuracy, F1-score) on the percentual scale. In principle, our framework using inflated Beta distribution can accommodate Cohen's Kappa via min–max normalisation that would bring its values to [0,1] range. For each metric and algorithm, the predictor functions of each distributional parameter (μ, σ, ν, τ) were identified using a forward stepwise regression procedure with Generalized Akaike Information Criterion (GAIC) in the `gamlss` R package. The stepwise procedure starts from an intercept-only model and iteratively adds the most significant predictor, if available. For the functional form of predictors, we considered the binary gender-male indicator, linear terms of AHI and PLMI, and a P-spline[155] age-term with 3 degrees of freedom (df). Whereas the binary/linear terms enable quantification of a uniform change (increase/decrease) in outcome (performance) due to a unit change in X (i.e., gender, AHI, PLMI), the spline age-term with df = 3 enables data-driven estimation of non-linear effects, with up to two inflection points (local extremes). The rationale behind the choice of linear effects for AHI and PLMI is that these measures of breathing and movement disturbances are expected to proportionally degrade performance and increase its variability. In contrast, age was modelled using splines to account for potential nonlinear changes in sleep structure at the biological level [18], [78], [79], which we anticipated would manifest as non-uniform shifts in the performance distribution.

Table 4.2 presents the estimated predictors of each performance metric and the two sleep-scoring algorithms. The spline term of age significantly influenced the location (μ) and scale (σ) in all cases, indicating a nonlinear relationship across age and both, expectation and variability of U-Sleep and YASA performance metrics. A detailed view of these effects can be seen in Supplementary Figures A.1-A.4. AHI and PLMI were significantly associated with lower expected performance (decreasing μ) and higher variability (increasing σ) across both models, reinforcing the challenge of scoring individuals with frequent breathing arrests and movement disturbances. Gender effects were minimal, with a significant impact observed only in U-Sleep's location parameter, suggesting potential gender-related disparities beyond age and AHI/PLMI effects. Notably, YASA showed no significant gender-based performance differences. Finally, neither zero-inflation (ν) nor one-inflation (τ) showed significant associations with any bias-inducing factors. However, this lack of significance likely results from the scarcity of extreme performance cases (<5 such PSGs per scenario). More details on individual effects can be obtained from the supplementary app, specifically `Model Summary` and `Partial Effects` tabs.

Performance Across Demographic and Clinical Subgroups Figure 4.1 and Supplementary Figure A.5 depict the expected distribution of subject-specific F1-scores and accuracy for U-Sleep. Parts (i)-(iii) and (iv)-(vi) demonstrate the marginal distribution of age, AHI, and PLMI, for males and females, respectively. For both metrics and genders, the highest performance with the least variability (i.e., the narrowest quantile spread) is observed in mid-20s females with no apnea- or movement-related events (AHI = PLMI = 0). In this group, the median F1-score and accuracy reach 81.7% and 86.5%, respectively, with 1-99% quantiles spanning 63.3–93.7% (F1-score) and 70–96.1% (accuracy). In all quantiles, males reach slightly lower performance, with difference <1%, reflecting the significant μ term in Table 4.2. Performance declines notably in pediatric subjects, and variability increases at both younger and older ages (cf. Supplementary Figures A.1-A.2). For instance, newborn and 75-year-old females exhibit median F1-scores of 59.2% and 74.3%, with corresponding 1-99% ranges of 32.3–82.9% and 46.6–92.7%. The accuracy (Supplementary Figure A.5) follows a similar trend, as it is closely related to the F1-score. These systematic and statistically significant distributional age-related shifts in algorithmic performance can be attributed to several factors. *First*, imbalances and under-representation of different age groups in U-Sleep training data, e.g., only 1 in 14 databases is children-based [60]. *Second*, the algorithm

Table 4.2: Significant predictors in performance quantification models for U-Sleep and YASA.

Performance metric	Sleep-scoring algorithm	Performance-model's parameter	Intercept	Performance-models' terms pb(Age)	AHI	PLMI	Gender
Accuracy	U-Sleep	$logit(\mu)$	1.629	*	-0.008	-0.003	-0.044
		$logit(\sigma)$	-1.636	*	0.004	0.003	-
		$log(\nu)$	-21.538	-	-	-	-
		$log(\tau)$	-7.191	-	-	-	-
	YASA	$logit(\mu)$	1.345	*	-0.010	-0.002	-
		$logit(\sigma)$	-1.449	*	0.006	0.003	-
		$log(\nu)$	21.538	-	-	-	-
		$log(\tau)$	-7.597	-	-	-	-
F1-score	U-Sleep	$logit(\mu)$	1.286	*	-0.006	-0.003	-0.043
		$logit(\sigma)$	-1.522	*	0.003	0.003	-
		$log(\nu)$	-21.538	-	-	-	-
		$log(\tau)$	-7.191	-	-	-	-
	YASA	$logit(\mu)$	0.900	*	-0.008	-0.003	-
		$logit(\sigma)$	-1.281	*	0.004	0.003	-
		$log(\nu)$	-20.538	-	-	-	-
		$log(\tau)$	-7.597	-	-	-	-

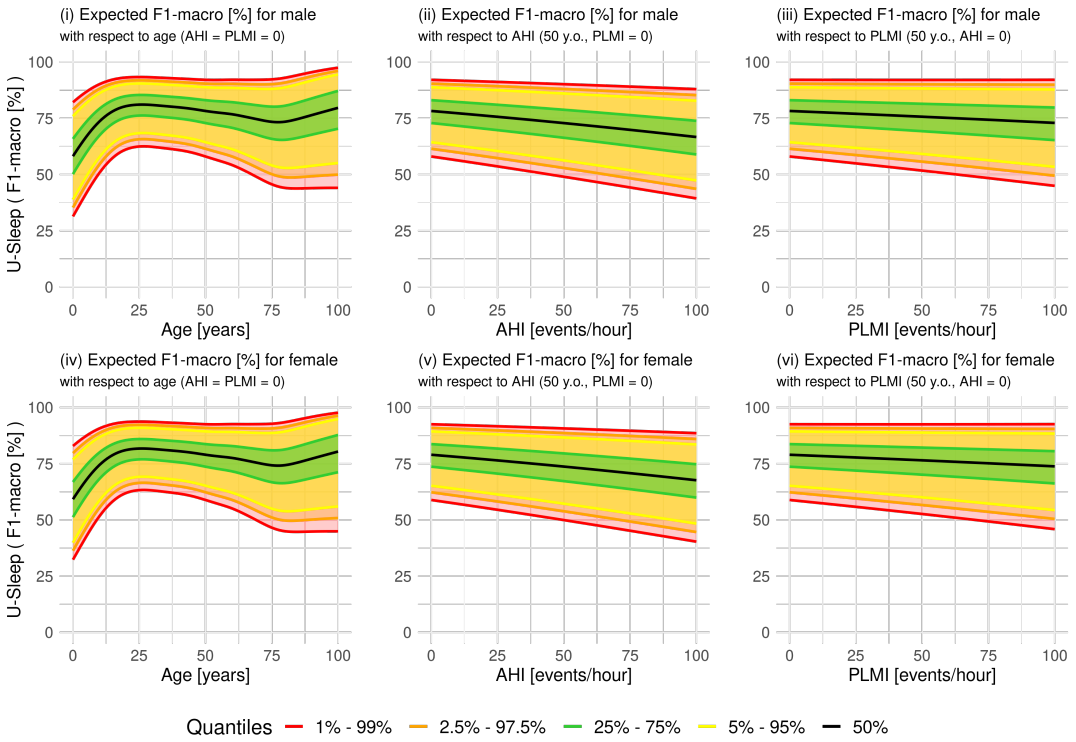
Notes: Significant predictors for the zero-and-ones-inflated Beta distribution GAMLSS model used for performance quantification of accuracy and macro F1-score in U-Sleep and YASA sleep-scoring algorithms. The table reports parameter estimates for the model's location (μ), scale (σ), zero-inflation (ν), and one-inflation (τ) terms. Predictors include the intercept, age (modelled as a spline term, pb(Age)), AHI, PLMI, and gender. "*" indicates a significant spline term, while "-" denotes non-significance. The spline term for age, pb(Age), which cannot be summarized by a single coefficient, can be further explored through the interactive application supplementing our study.

may struggle to learn clinically established age-related EEG changes [79], as evidenced by increasing performance variability and decreasing accuracy in older subjects. *Third*, since U-Sleep does not incorporate age as an input [59], it cannot capture age-EEG interactions. *Lastly*, AASM guidelines define distinct sleep-scoring rules for pediatric cases [8]. The combination of missing age information and the under-representation of children in the training data makes pediatric sleep staging particularly challenging. In addition to the age-related performance shifts, panels (ii)-(iii) and (v)-(vi) demonstrate the AHI-PLMI effects for males and females, respectively. For both AHI and PLMI, the performance metrics significantly decrease and their variability increases, particularly with AHI. As shown in the Figure 4.1, the median and 1-99% range in F1-score for 50-year-old female with AHI = PLMI = 0 decreases from 79% (58.9-92.6) to 73.7% (50-90.7) if AHI = 50 and PLMI = 0, and to 76.5% (52.6-92.5) if AHI = 0 and PLMI = 50. These changes are likely related to movement artefacts possibly introduced by both AHI and PLMI, but also to the altered sleep patterns, such as increased sleep stage transitions and reduced sleep efficiency, that all correlate also with age and may pose challenges to both, sleep-scoring algorithms and also human-scorers, who tend to reach lower agreement for PSGs of sleep-disordered subjects [67]–[69].

Similar trends are observed in Supplementary Figures A.6 and A.7, and A.3-A.4, which illustrate the distributions of F1-score and accuracy, and related age effects, for YASA, respectively. Unlike U-Sleep, none of the distributional parameters of YASA's performance metrics were significantly impacted by gender (cf. Table 4.2), indicating the absence of gender bias in YASA predictions. Consistently with U-Sleep, maximal performance with minimal variability, of 74% (50.4-90.8) and 82.2% (61.9-94.7) in median and 1-99% range of F1-score and accuracy, respectively, is reached for subjects in their mid-20s. The trend of declining performance, down to 55% (39.2-80.8) in F1-score in newborns and 66.6% (38.5-88.6) in 75-year-old subjects, is consistent with U-Sleep, reflecting the same underlying challenges. YASA's performance exhibits a more pronounced decline with AHI, as indicated by the estimated μ effects from Table 4.2, which are larger in magnitude compared to U-Sleep (e.g., -0.008 vs. -0.01 for U-Sleep and YASA accuracies, respectively). Similarly, YASA's performance variability tends to be larger, as reflected by greater σ effects. In contrast, PLMI had comparable effects on both μ and σ in both algorithms. As shown in Supplementary Figure A.6, for a 50-year-old subject with AHI = PLMI = 0, the median and 1–99% range in F1-score decreases from 71.8% (45.9–90.4) to 63.1% (32.5–87.7) if AHI = 50, PLMI = 0, and to 68.5% (39.3–90.1) if AHI = 0, PLMI = 50.

In summary, YASA exhibits about 5% lower performance than U-Sleep and follows a very similar trend concerning subjects' age. In addition, YASA seems to be more sensitive

Figure 4.1: Expected distribution of subject-specific F1-score for U-Sleep across demographic and clinical subgroups.



Notes: Expected distribution of the subject-specific macro F1-score based on the zero-and-ones-inflated Beta performance model for U-Sleep predictions, stratified by gender (top row: males, bottom row: females). The graphs display the estimated distribution as a function of bias-inducing variables (age, AHI, and PLMI) on the horizontal axis. The solid black line represents the median, while the shaded areas correspond to different percentile ranges: 25-75% in green, 5-95% in yellow, 2.5-97.5% in orange, and 1-99% in red, illustrating the expected performance variability across subjects' characteristics.

to breath-disruption-related artefacts and sleep alterations, as its performance declines more with AHI. These results suggest that the compared version of deep-learning-based U-Sleep has superior sleep-scoring capability than YASA. However, as both were trained on different amounts of data, we cannot conclude, whether U-Sleep superiority is attributable to a more complex model's architecture or a larger set of training data. Next, unlike YASA, U-Sleep predictions exhibit a small but statistically significant gender bias. Finally, arbitrary results quantifying performance distribution of both algorithms and specified age, gender, AHI, and PLMI, including their marginal effects, can be explored in our interactive app.

4.3.4 Algorithmic bias concerning clinical markers

While the previous section evaluated the overall predictive performance of the algorithms in sleep stage classification, this section explores whether algorithmic predictions accurately preserve clinically relevant PSG markers and whether errors in derivations of these markers are random or systematically influenced by bias-inducing variables X.

Descriptive statistics of PSG markers based on physicians' and algorithms' sleep-scoring: Supplementary Table A.1 summarizes the statistical characteristics (mean, SD, quantiles, min, max) of PSG markers derived from reference *physicians' scoring* and (U-Sleep, YASA) *algorithmic predictions*, based on our BSWR dataset. Additionally, the results of Wilcoxon signed-rank tests, assessing whether differences between algorithmic and physician-based values are symmetrically distributed around zero, are presented. The values of PSG markers span their full possible ranges: the (sleep, REM) latencies vary from 0 minutes to hours, hypnograms include cases with no recorded sleep (i.e., total sleep time and sleep efficiency), and sleep-stage percentages exhibit broad variability. For example, the mean (SD), min-max range of W% was 19.9 (14.9), 0-100 for physician-based scoring, whereas 18.5 (14.4), 0-100

and 28.7 (17.2), 1.9-100 based on derivation from U-Sleep and YASA predictions, suggesting possible U-Sleep under-estimation and YASA over-estimation of wakefulness percentages. Except for sleep cycle counts for YASA, Wilcoxon-based comparisons revealed significant deviations from zero in error distributions for all PSG markers across both sleep-scoring algorithms, indicating systematic shifts in their predictions.

Expected clinical trends and their potential impact on bias interpretation: Established clinical literature describes several age-, gender-, and pathology-related trends in hypnogram-derived markers that are important to consider when interpreting potential biases. For example, N3 sleep is known to decline with age [17], [18], [156], which could amplify relative errors or lead to proportional overestimation in subjects with already low N3 duration. Total sleep time (TST) tends to decrease and wake after sleep onset (WASO) increases with age [17], [156], [157], which may affect the accuracy of TST and WASO predictions in older adults. Similarly, AHI and PLMI—both of which increase with age and are more common in males and older subjects [158], [159]—are linked to more fragmented sleep, increased awakenings, N1, and N2, and reductions in REM and N3 stages [18]. These factors are also known to elevate inter-scorer disagreement, particularly in clinically complex subjects [67]–[69], [88], which may increase the uncertainty of the reference labels used for evaluation. As a result, there is a potential for both overestimation and underestimation of certain markers by automated algorithms. Moreover, elevated label variability in these subgroups could lead to an inflation of the observed bias spread, even in the absence of systematic over- or under-prediction. By explicitly including age, gender, AHI, and PLMI as covariates for both location (μ) and scale (σ) distributional parameters in our proposed bias-quantification framework, it is designed to account for these expected influences.

Descriptive statistics of algorithmic errors in PSG markers Supplementary Table A.2 presents the characteristics of raw errors in PSG markers, derived from algorithm-predicted and physicians' sleep-scoring. A negative mean or median (Q50) raw error suggests that the model underestimates a given marker, meaning the predicted value is, on average, lower than the reference one, in the majority of evaluated PSGs. For example, the percentage of wakefulness after sleep onset (W%) is on-average underestimated by 1.3% by U-Sleep, whereas YASA overestimates it by 8.8%. Median errors, which are less sensitive to outliers, confirm this trend, with values of -0.6 and 6.2 for U-Sleep and YASA, respectively. The smaller magnitude of the observed errors for U-Sleep may be attributable to its generally higher sleep-scoring performance identified above. Further, the systematic under- and over-estimation of W% by the two models propagate to related sleep metrics: U-Sleep underestimates sleep latency, WASO, and the number of awakenings per hour, while YASA overestimates them. Conversely, U-Sleep overestimates total sleep time and sleep efficiency, whereas YASA underestimates them. Further, Table 4.3 describes absolute errors, which disregard under- and overestimation, and quantify the overall magnitude of algorithmic errors. For W%, U-Sleep has a mean (SD) absolute error of 3.2 (4.9), indicating lower bias and its spread compared to YASA's 9.1 (9.5). Additionally, the best-performing 10% of subjects (Q10) had an absolute error below 0.3% for U-Sleep and 1.8% for YASA, suggesting a six-fold larger error magnitude in YASA's best cases. In contrast, the worst 10% (Q90) exhibited absolute errors of at least 7.7% for U-Sleep and 19.5% for YASA. Table 4.3 enables corresponding interpretation for all of the PSG markers listed. The merit of using absolute errors and their quantiles lies in their intuitive interpretability—they provide a clear summary of the error distribution across the dataset. For example, a Q50 (median absolute error) of 7.0 minutes for U-Sleep and 26.0 minutes for YASA in total sleep time (TST) suggests that, for half of the evaluated PSGs, the predicted TST deviated from the reference by less than ±7 and ±26 minutes, respectively. In addition, Table 4.3 presents the Spearman correlation of absolute errors with age, AHI, and PLMI, alongside gender-based error differences assessed via the Wilcoxon test. Except for sleep cycle count (for both algorithms) and the correlation between age and awakenings per hour (for YASA), absolute errors in all PSG markers and both models show significant correlations with age, AHI, and PLMI. The signs of these correlations were shared among the age, AHI, and PLMI, likely due to their joint (positive) association with age (cf. Table 4.3). Among the significant associations, the absolute errors, and hence the error-bounds, tend to increase with age, AHI, and PLMI for most PSG markers in both

models. However, an exception is observed in N3 and REM%, which are clinically known to decrease with age [18], and likely therefore the error magnitude tends to be smaller (also in age-positively correlated AHI and PLMI). Regarding gender differences, out of 13 evaluated PSG markers, U-Sleep exhibited 10 significant gender-based differences in absolute error, with 8 overestimations and 2 underestimations in males. In comparison, YASA showed 6 such differences, with 4 and 2 over- and underestimation in males.

Table 4.3: Summary of absolute errors in PSG markers derived from U-Sleep and YASA compared to physician scoring.

Absolute Error	Algorithm	Mean	SD	Q10	Q25	Q50	Q75	Q90	Min	Max	ρ(Age)	ρ(AHI)	ρ(PLMI)	M - F
Sleep Latency	U-Sleep	5.5	13.8	0.0	0.5	1.5	4.5	13.5	0.0	271.0	0.09	0.09	0.09	-0.50
[*minutes*]	YASA	9.8	18.5	0.5	1.0	4.0	10.5	23.5	0.0	266.5	0.17	0.13	0.10	-1.00
REM Latency	U-Sleep	30.2	62.2	0.5	1.0	3.0	16.0	108.0	0.0	608.0	0.04	0.06	0.05	0.00
[*minutes*]	YASA	45.5	67.2	1.0	2.5	10.0	70.5	144.3	0.0	411.5	0.09	0.10	0.09	0.00
Total Sleep Time	U-Sleep	13.5	21.2	1.0	3.0	7.0	15.5	31.5	0.0	353.5	0.17	0.18	0.13	1.50
[*minutes*]	YASA	38.3	41.3	7.5	14.5	26.0	47.0	81.5	0.0	532.5	0.13	0.22	0.14	1.00
WASO	U-Sleep	12.8	20.6	1.0	2.5	6.0	14.5	30.5	0.0	317.5	0.18	0.19	0.14	1.50
[*minutes*]	YASA	31.8	36.0	4.5	10.5	21.0	40.0	69.5	0.0	426.5	0.09	0.19	0.12	1.50
Sleep Cycles	U-Sleep	0.3	0.6	0.0	0.0	0.0	0.5	1.0	0.0	8.0	0.01	0.01	-0.02	0.00
[N]	YASA	0.5	0.7	0.0	0.0	0.0	1.0	1.0	0.0	10.5	-0.02	0.02	-0.01	0.00
Transitions	U-Sleep	7.2	5.1	1.4	3.4	6.4	10.1	13.9	0.0	39.8	0.16	0.22	0.04	0.30
[$N/hour$]	YASA	5.9	4.7	0.9	2.4	4.8	8.5	11.9	0.0	38.2	0.20	0.23	0.06	0.54
Awakenings	U-Sleep	1.0	1.2	0.1	0.3	0.6	1.2	2.1	0.0	17.1	0.13	0.20	0.07	0.15
[$N/hour$]	YASA	1.9	1.8	0.2	0.6	1.4	2.7	4.3	0.0	15.6	0.01	0.14	0.04	-0.05
Sleep Efficiency	U-Sleep	3.2	4.9	0.3	0.7	1.7	3.7	7.7	0.0	66.3	0.20	0.20	0.14)	0.47
[%]	YASA	9.1	9.5	1.8	3.4	6.3	11.4	19.5	0.0	99.9	0.17	0.24	0.14	0.48
W	U-Sleep	3.2	4.9	0.3	0.7	1.7	3.7	7.7	0.0	66.3	0.20	0.20	0.14	0.47
[%]	YASA	9.1	9.5	1.8	3.4	6.3	11.4	19.5	0.0	99.9	0.17	0.24	0.14	0.48
N1	U-Sleep	6.6	6.6	0.7	2.1	4.7	9.0	14.8	0.0	61.5	0.27	0.30	0.14	1.37
[%]	YASA	11.3	9.5	2.2	4.6	8.8	15.1	23.5	0.0	77.4	0.39	0.46	0.23	2.61
N2	U-Sleep	9.7	8.4	1.5	3.7	7.6	13.2	20.1	0.0	78.3	0.12	0.24	0.10	0.63
[%]	YASA	7.3	6.8	0.9	2.5	5.4	10.0	16.0	0.0	78.1	0.25	0.28	0.11	0.90
N3	U-Sleep	4.7	5.0	0.4	1.3	3.3	6.5	11.0	0.0	75.3	-0.09	-0.07	-0.03	-0.53
[%]	YASA	4.3	4.9	0.3	1.1	2.7	5.7	10.3	0.0	75.3	-0.09	-0.09	-0.04	-0.29
REM	U-Sleep	2.0	2.5	0.1	0.5	1.2	2.6	4.6	0.0	37.6	-0.10	-0.08	-0.07	-0.11
[%]	YASA	2.7	3.1	0.3	0.8	1.8	3.6	6.0	0.0	39.8	-0.07	-0.04	-0.03	-0.04

Notes: Summary of absolute errors in sleep metrics, derived from hypnograms predicted by U-Sleep and YASA compared to those derived from physician-scored hypnograms. The table presents the mean absolute error, standard deviation (SD), (10, 25, 50, 75, 90)%-quantiles (Q10, Q25, Q50, Q75, Q90), minimum (Min), and maximum (Max) values. Spearman correlations (ρ) evaluate the association of absolute errors and bias-inducing factors (age, AHI, and PLMI) across various sleep metrics. The table also includes the median difference in absolute errors between males and females (M - F), assessed by the Wilcoxon test. Significant associations and differences are highlighted as p-value < 0.05 , 0.01 , and 0.001 , respectively.

Bias-Quantification Models The descriptive statistics of error distributions in PSG markers provided an overview of the general trends of algorithmic biases, and correlations assessed their pairwise associations. Despite the level of detail, conclusions about potential biases should not rely solely on descriptive statistics. Measures such as averages or quantiles do not account for the uneven distribution of the population, for instance, variations in age or clinical characteristics (AHI, PLMI), nor do they reflect how these factors individually influence the distribution of prediction errors. To systematically quantify biases while accounting for all bias-inducing variables X (age, AHI, PLMI, and gender) simultaneously, we modelled the differences between algorithm-predicted and physician-based reference PSG markers using the generalized normal distribution from Eq. 4.3. Each distributional parameter—location (μ) and scale (σ)—was modelled with unique predictors. As in the performance models, predictor functions for each distributional parameter were selected via a forward stepwise regression procedure using the Generalized Akaike Information Criterion (GAIC). Using the same rationals as for the performance quantification models, the candidate predictors included gender (male) indicator, linear terms of AHI and PLMI, and the cubic spline of age. Table 4.4 summarizes the significant predictors (X) for location (μ) and scale (σ) in each marker-specific bias-quantification (BQ) model. The effects of each predictor in these models can be interpreted as adjusted for all other included variables. In some cases, certain predictors were not included, because their inclusion did not improve the GAIC, meaning, their presence would increase model complexity without enhancing goodness-of-fit.

A key finding is that for both U-Sleep and YASA, bias distribution of all PSG markers was significantly influenced by the spline term of age in both location and scale parameters. This suggests that the magnitude and variability of errors in both algorithms systematically and non-linearly differ across ages. Several clinical and technical factors likely contribute to this result, paralleling those discussed in performance quantification: (i) AASM guidelines define different scoring rules for pediatric and adult subjects [8], (ii) sleep architecture differs across age groups (e.g., variations in sleep-stage percentages) [18], [78], [139], (iii) raw EEG signals evolve with age and may carry age-related artifacts [79]. Since neither U-Sleep nor YASA incorporates age as its input [58], [59], these age-related variations are likely ignored in both models. Consequently, the algorithms may underfit the AASM guidelines due to omitted variable bias (as age defines different scoring rules for pediatric vs adult subjects), or overfit specific age-related EEG patterns [79] correlated with sleep stages, rather than learning scoring rules directly. It is important to mention that Fiorillo et al. [60] conducted extensive computational experiments to assess whether the fine-tuning of U-Sleep on specific age groups could enhance its performance in age-matched test data. However, applying *sandwich batch normalization* (SaBN)—a rolling standardization technique similar to z-score normalization conditioned on age—did not yield significant improvements [60], as it may have failed to capture the age-related EEG alterations that are manifested primarily in spectral (frequency) domain [79]. A more straightforward approach could involve applying SaBN selectively to specific frequency bands incorporating age as an additional input to the model, or using age-stratified sampling during the training.

Further, both AHI and PLMI were associated with the increased error variability (σ) and seem to worsen the baseline bias (μ), meaning, they typically increase the magnitude of the bias in the direction identified for specific ages. The only exceptions were the σ in the number of sleep-cycles (not influenced by AHI in U-Sleep; decreasing with PLMI in both algorithms) and the N3% (decreasing with AHI in both algorithms; decreasing with PLMI in YASA). These exceptions may likely be related to reductions in N3 and REM sleep evidenced in subjects with movement- and breath-related sleep disorders [18], [139], [151].

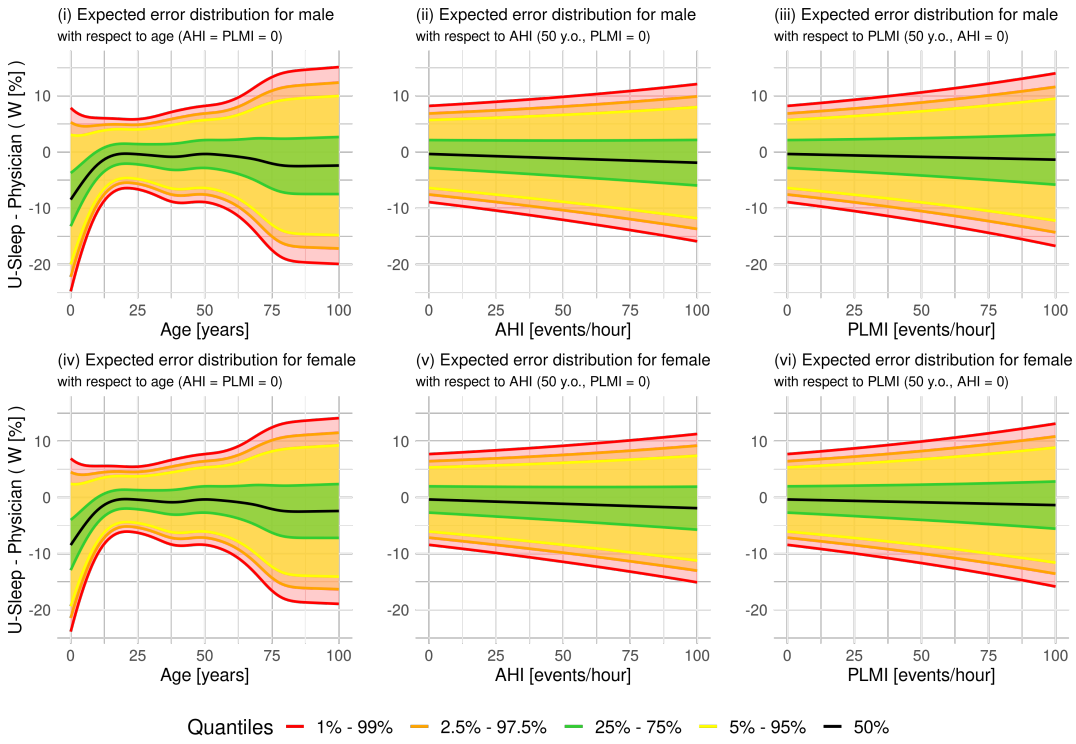
Out of 13 PSG markers, gender significantly influenced bias in 7–8 (μ-σ) parameters for U-Sleep and 10–9 respective parameters for YASA, suggesting that YASA exhibits more gender-related biases than U-Sleep after controlling for age, AHI, and PLMI. This also suggests that the larger count in significant gender differences in descriptive error statistics for U-Sleep (cf. Table 4.3), may rather be attributed to the age- rather than gender-related bias, as males our study dataset are significantly older (cf. Table 4.1). After adjusting for all bias-inducing variables simultaneously, U-Sleep seems to be less affected by gender-bias compared to YASA. Most prominent gender-related biases included the male-specific effect of 1.29 in μ of N3% bias for U-Sleep, -1.32 minutes in μ in sleep latency for YASA, and 5.48

minutes in μ of REM latency error for U-Sleep.

Based on the estimated bias models from Table 4.4, Figure 4.2 illustrates the expected bias distribution for W% in U-Sleep. Sub-figures (i) and (iv) show the expected bias for males and females under an optimistic scenario where AHI = PLMI = 0. Since gender does not significantly affect μ, the trend remains consistent between males and females; however, variance is greater in males due to significant σ term (cf. Table 4.4). W% is underestimated in pediatric cases (age <18 years), likely due to the already discussed differences in AASM scoring rules. Bias is minimal with the highest precision (lowest variance) in early adulthood (20s), followed by increasing underestimation and variability with age. The partial effects of age on μ and σ of W% bias in U-Sleep is provided in Supplementary Figure A.8 and can be compared with Supplementary Figure A.9 for YASA, which overestimates the W%. Finally, both AHI and PLMI from panels (ii, v) and (iii, vi) of Figure 4.2 amplify underestimation and increase variance, as indicated by a decreasing median (Q50) trend and widening 1–99% quantile range.

Corresponding results quantifying bias distribution for all evaluated PSG markers and arbitrary age, gender, AHI, and PLMI for both U-Sleep and YASA, including their marginal effects, can be explored in our interactive app. Supplementary Figures A.1-A.22 and A.1-**??** provide a detailed view of the bias distribution regarding individual clinical markers derived from YASA and U-Sleep predictions, respectively, including their dependencies on age, gender, AHI, and PLMI.

Figure 4.2: Expected distribution of the bias in the wakefulness percentage after sleep onset (W, %) for U-Sleep predictions.



Notes: Expected distribution of the bias in the percentage of wakefulness (W%) based on the generalized normal distribution for U-Sleep predictions, stratified by gender (top row: males, bottom row: females). The graphs display the estimated distribution as a function of bias-inducing variables (age, AHI, and PLMI) on the horizontal axis. The solid black line represents the median, while the shaded areas correspond to different percentile ranges: 25-75% in green, 5-95% in yellow, 2.5-97.5% in orange, and 1-99% in red, illustrating the expected performance variability across subjects' characteristics.

4.3.5 Utilizing biased predictions for diagnostic purposes

The derived PSG markers are in clinical practice widely used for diagnostics. In previous sections, we highlighted potential biases in sleep-scoring algorithms when predicting these markers. This section evaluates whether simple machine learning (ML) classifiers trained on

Table 4.4: Significant predictors of bias quantification models for PSG markers in U-Sleep and YASA.

PSG parameter	Sleep-scoring algorithm	Bias-model's parameter	Bias-models' terms: Intercept	pb(Age)	AHI	PLMI	Gender
Sleep Latency [*minutes*]	U-Sleep	μ	-1.628	*	-0.020	-0.038	0.768
		$log(\sigma)$	2.523	*	0.003	0.005	-0.107
	YASA	μ	7.027	*	0.061	0.079	-1.321
		$log(\sigma)$	2.518	*	0.004	0.008	-
REM Latency [*minutes*]	U-Sleep	μ	-8.839	*	-	0.115	5.477
		$log(\sigma)$	4.023	*	0.004	0.002	-0.168
	YASA	μ	-21.570	*	-	-	-
		$log(\sigma)$	4.161	*	0.004	0.003	-0.079
Total Sleep Time [*minutes*]	U-Sleep	μ	1.027	*	0.055	0.038	-
		$log(\sigma)$	2.666	*	0.004	0.006	0.045
	YASA	μ	-28.886	*	-0.443	-0.153	3.203
		$log(\sigma)$	3.490	*	0.008	0.004	-0.125
WASO [*minutes*]	U-Sleep	μ	0.110	*	-0.036	-	-0.377
		$log(\sigma)$	2.709	*	0.005	0.005	-
	YASA	μ	21.504	*	0.388	0.078	-1.797
		$log(\sigma)$	3.386	*	0.008	0.004	-0.051
Sleep Cycles [N]	U-Sleep	μ	0.243	*	-0.001	-0.001	-
		$log(\sigma)$	-0.560	*	-	-0.002	0.075
	YASA	μ	0.083	*	-	-	-
		$log(\sigma)$	-0.410	*	0.003	-0.001	0.057
Transitions [$N/hour$]	U-Sleep	μ	-6.362	*	-0.048	0.001	-
		$log(\sigma)$	1.499	*	0.009	0.002	-
	YASA	μ	-3.826	*	-0.048	-	-
		$log(\sigma)$	1.589	*	0.007	0.002	-
Awakenings [$N/hour$]	U-Sleep	μ	0.071	*	-0.010	0.001	-
		$log(\sigma)$	-0.106	*	0.011	0.002	0.114
	YASA	μ	1.347	*	0.012	0.004	-0.255
		$log(\sigma)$	0.552	*	0.009	0.002	-0.054
Sleep Efficiency [%]	U-Sleep	μ	0.359	*	0.015	0.010	-
		$log(\sigma)$	1.244	*	0.005	0.006	0.058
	YASA	μ	-6.892	*	-0.109	-0.036	0.619
		$log(\sigma)$	1.977	*	0.008	0.004	-0.072
W [%]	U-Sleep	μ	-0.359	*	-0.015	-0.01	-
		$log(\sigma)$	1.244	*	0.005	0.006	0.058
	YASA	μ	6.892	*	0.109	0.036	-0.619
		$log(\sigma)$	1.977	*	0.008	0.004	-0.072
N1 [%]	U-Sleep	μ	-3.634	*	-0.086	-0.013	-0.065
		$log(\sigma)$	1.558	*	0.010	0.004	0.101
	YASA	μ	-6.557	*	-0.205	-0.054	-0.298
		$log(\sigma)$	1.683	*	0.010	0.003	0.081
N2 [%]	U-Sleep	μ	6.231	*	0.111	0.030	-0.737
		$log(\sigma)$	1.831	*	0.008	0.002	-
	YASA	μ	1.193	*	0.094	0.020	0.526
		$log(\sigma)$	1.845	*	0.009	0.002	-
N3 [%]	U-Sleep	μ	-3.794	*	-	-	1.294
		$log(\sigma)$	1.832	*	-0.003	0.001	-0.074
	YASA	μ	-2.791	*	0.015	-	0.732
		$log(\sigma)$	1.872	*	-0.003	-0.001	-0.158
REM [%]	U-Sleep	μ	1.597	*	-0.008	-0.007	-0.293
		$log(\sigma)$	0.866	*	0.002	0.001	-
	YASA	μ	0.765	*	-0.009	-	-
		$log(\sigma)$	1.218	*	0.002	0.002	-

Notes: Significant predictors for the extended normal distribution GAMLSS model used for bias quantification of U-Sleep and YASA in predicting key clinical markers derived from PSG data. The table reports parameter estimates for the model's location (μ) and scale (σ), alongside the significance of bias-model terms, including the intercept, age (as a spline term, pb(Age)), AHI, PLMI, and gender. "*" indicates a significant spline term, while "-" denotes non-significance. The bias in each PSG-derived marker was assessed by comparing derivations based on hypnograms predicted by (U-Sleep/YASA) with those obtained from physicians.

PSG markers derived from physician-scored and algorithm-predicted hypnograms can mitigate these biases and effectively identify the presence of Obstructive Sleep Apnea (OSA). OSA was chosen as the target condition due to its high prevalence—affecting approximately 17% of the general population [149]—its role as a significant risk factor for overall mortality and cardiovascular events [150], and its well-documented impact on sleep patterns [18], [139], [151]. For this analysis, we identified a subset of 678 PSG recordings from 641 unique

subjects (69% males), with a mean (SD) age of 52.1 (15.7) years, where a conclusive diagnosis of either OSA—616 PSGs from 579 subjects aged 53.9 (14.3), 72.7% males—or healthy control—62 PSGs from 62 subjects aged 34.9 (18), 40.3% males—was available. This subset passed the exclusion criteria of failing to fall asleep (i.e., total sleep time = 0), application of respiratory therapy (e.g., CPAP), or if excessive light exposure affected more than 5% of the PSG-study duration. As predictors, we employed 29 hypnogram-derived PSG markers:

- Minutes $[mins]$ of (sleep, W, N3, REM)-latencies, Wake-After-Sleep-Onset (WASO), Total-Sleep-Time (TST), mean- and maximum-durations of (W, N1, N2, N3, REM), totalling 16 variables;
- Percentages $[\%]$ of (W, N1, N2, N3, REM)-stages after sleep onset, and sleep efficiency, totalling 6 variables;
- Counts $[N]$ and rates $[N/hour]$ of sleep-stage-transitions and awakenings, totalling 4 variables;
- Count $[N]$ of NREM-REM cycles different bout durations (0,3, and 10 minutes), totalling 3 variables,

and their interactions with age (in decades) and gender (binary indicator: 1 = male, 0 = female). This resulted in a total of $87 = 3 \times 29$ hypnogram-based predictors, supplemented by age and gender. These predictors were computed based on both physician-scored and algorithm-predicted hypnograms (U-Sleep and YASA).

We evaluated five ML classifiers: Linear Discriminant Analysis (LDA), LASSO and Ridge logistic regression, Random Forest (RF), and K-Nearest-Neighbors (KNN). To estimate uncertainty in performance metrics, we applied a cross-validation (CV) strategy, and partitioned PSG recordings into five approximately equal-sized subject-wise splits (folds/groups) using anticlustering [160], ensuring balanced distributions of age, gender, and OSA prevalence across folds. Model hyperparameters, such as the regularization parameter λ for LASSO and Ridge, and the number of neighbors K for KNN, were optimized using an inner three-fold CV applied to the training data (i.e., 4 out of 5 folds in each CV run).

Table 4.5 presents the mean (SD) of performance metrics for ML classifiers trained on PSG markers derived from physician-scored and algorithm-predicted hypnograms (U-Sleep, YASA). The results indicate that classification performance using algorithm-derived markers was comparable to that based on physician-derived markers, as the 95% confidence intervals (mean $\pm$ 1.96$\times$SD) of the physician-based scenario overlapped with those of both U-Sleep and YASA across all methods and metrics. Since the performance intervals overlap, no superiority or inferiority can be inferred for either algorithm.

Despite systematic biases in algorithm-predicted PSG markers (see Table 4.4), ML classifiers effectively adapted to these shifts, achieving comparable AUROC values, maintaining strong classification performance (AUROC > 80% across all scenarios). Accuracy, sensitivity, and specificity were also consistent across sleep-scoring sources, although slightly lower sensitivity was observed in some methods (e.g., with KNN), likely due to probability calibration issues—evidenced by corresponding increases in specificity. These findings suggest that, despite inherent biases, algorithm-derived PSG markers can serve as informative inputs for ML-based diagnostics, with classifiers effectively adapting to systematic deviations in algorithm-predicted PSG markers, preserving diagnostic utility for OSA detection.

4.4 Discussion

With the advancement of AI and data-driven algorithmic solutions, there is a growing demand for validation approaches that address not only performance but also generalization and compliance with regulatory standards. This is particularly critical in healthcare, where AI holds tremendous potential to improve efficiency and increase accessibility to medical care. However, these potential benefits come with the risk of unfairness, especially concerning sensitive attributes such as age, gender, or specific clinical characteristics. When

Table 4.5: Performance comparison of machine learning classifiers trained on physician- and algorithm-derived PSG markers for OSA detection.

Method	Sleep-Scoring	AUROC	Accuracy	Sensitivity	Specificity
LDA	Physicians	85.5 (2.9)	91.4 (1.8)	47.1 (6.4)	95.9 (1.1)
	U-Sleep	84.7 (4.6)	91.4 (0.9)	49.5 (7.7)	95.6 (1.6)
	YASA	83.4 (2.6)	89.4 (1.0)	37.5 (8.1)	94.7 (2.1)
LASSO	Physicians	87.0 (1.8)	76.3 (2.6)	77.3 (6.4)	76.2 (3.2)
	U-Sleep	89.4 (1.6)	78.5 (1.7)	81.3 (9.1)	78.3 (2.1)
	YASA	84.5 (2.7)	75.8 (3.7)	77.3 (8.7)	75.7 (4.4)
Ridge	Physicians	86.8 (3.1)	77.5 (3.3)	79.3 (4.9)	77.3 (3.7)
	U-Sleep	88.7 (2.2)	80.7 (1.7)	82.5 (5.9)	80.5 (1.8)
	YASA	85.7 (2.3)	76.7 (3.6)	79.3 (4.9)	76.5 (3.9)
RF	Physicians	82.1 (3.7)	92.3 (1.4)	21.2 (8.0)	99.5 (0.4)
	U-Sleep	89.1 (2.6)	92.3 (1.3)	26.4 (13.5)	98.8 (1.0)
	YASA	82.6 (2.2)	92.2 (1.5)	19.2 (8.1)	99.5 (0.5)
KNN	Physicians	84.2 (3.8)	91.0 (1.7)	17.8 (7.7)	98.4 (1.6)
	U-Sleep	85.6 (3.7)	91.9 (1.2)	17.8 (2.9)	99.3 (0.4)
	YASA	83.5 (3.7)	90.7 (1.2)	13.2 (7.7)	98.5 (1.3)

Notes: Performance comparison of machine learning classifiers trained on PSG markers derived from physicians- and algorithm-based (U-Sleep, YASA) hypnograms to identify the Obstructive Sleep Apnea (OSA). The table presents the mean (standard deviation) of performance metrics: Area Under the Receiver Operating Characteristic Curve (AUROC), accuracy, sensitivity, and specificity, calculated using 5-fold cross-validation. The classifiers include Linear Discriminant Analysis (LDA), LASSO logistic regression, Ridge logistic regression, Random Forest (RF), and *K*-Nearest Neighbors (KNN).

predictive accuracy (e.g., for diagnosis, risk assessment, or clinical outcomes) is systematically shifted across these attributes or exhibits systematic dependencies on them, this is considered algorithmic bias [92].

Since most AI models in healthcare are trained on observational data, the risk of bias and reliance on spurious correlations often becomes a reality. This challenge has motivated regulatory frameworks (e.g., EU AI Act or MDR), which see the application of AI algorithms in healthcare as a high-risk area, mandate human oversight, and require fairness in their predictions. However, how to technically encounter fairness, or quantify its violation (bias) in practice, is still an open research field.

Standard validation practices typically involve reporting the average performance (e.g., accuracy, error rates) across subjects along with their variability, or the correlation of predicted and reference values. However, these conventional approaches do not provide insights into critical scenarios such as the minimum expected performance for different subpopulations (e.g., defined by restricted ranges of senstive attributes), its variability, or the potential dependence of both, predicted values or performance metrics, on external factors such as demographics or clinical profiles [134], [136], [142], [143], [146]. The relationship between algorithmic outcomes and sensitive attributes may be non-linear and exhibit varying degrees of variability, which current validation methods, such as the gold-standard Bland-Altman analysis used in clinical certification [134], fail to capture effectively [136].

Motivated by these limitations [136], [142], [143], [146], our study proposes a universal framework for quantifying algorithmic performance and bias, originally outlined in Bechny et al. (2024) [96], and applies it to the sleep medicine use-case of automatic sleep-scoring. The approach, based on the existing implementation of Generalized Additive Models for Location, Scale, and Shape (GAMLSS) [138], allows for flexible modeling of performance or bias distributions using a broad range of functional bases (including linear models, splines, or even neural networks) that quantify the relation between sensitive attributes and performance/bias distributional parameters [148], [155]. This enables detailed characterization of the performance and bias distribution, capturing the non-linear effects of sensitive attributes, and hypothesis testing to assess the presence of these relationship. In addition, upon quantification of the performance/bias distribution, expected quantiles may be reported with respect to the arbitrary values of included predictors (i.e., sensitive attributes), enabling critical assessments giving much broader idea of the expected algoritmic performance, compared to standard reporting.

We demonstrate this framework in the use case of automatic sleep-scoring, a particularly challenging task for AI algorithms due to inherent inter-rater variability among human experts (physicians) who provide the scoring labels used for training. Specifically, depending on metric, human scorers achieve an agreement-level of approximately 75-80% [14], [15], [66], [140], which introduces an inherent noise-level in the labels and defines a technical performance ceiling for AI-based solutions trained on sleep studies scored by multiple experts [141]. The necessity of human oversight (cf. EU AI Act) in this context is fully justified, as physicians remain responsible for clinical diagnostics that often rely on sleep-scoring—and even an optimal AI model can be expected to disagree with human scorers in approximately 20-25% of cases. We applied our approach to quantify the distribution of performance and errors for two widely recognized sleep-scoring algorithms: the deep-learning (DL)-based U-Sleep [59], [60] using raw biosignals (EEG, EOG) as its input, and the machine-learning (ML)-based YASA [58] using as its input derived statistical features. In this study, we demonstrated our framework from the perspective of end users (clinicians) and regulatory bodies, by evaluating two existing models trained on different data using a rich, unseen out-of-domain dataset.

To quantify performance metrics (e.g., accuracy, F1-score), we used a zero-and-one-inflated Beta distribution, which is defined on the 0–100% range and allows for the identification of cases of perfect or entirely incorrect classification through inflation parameters. Performance distributions were modelled while accounting for age, gender, AHI, and PLMI, enabling us to test whether predictive performance systematically varies with these factors. The deep-learning-based U-Sleep demonstrated higher expected performance and lower variability across demographic and clinical subgroups (AHI/PLMI), making it superior to the feature-based YASA. Both algorithms exhibited a significant non-linear effect of age, quantified by spline, on expected performance and its variability, with the most pronounced challenges occurring in pediatric cases. The primary reasons for this performance drop appear to be a combination of (i) the under-representation of pediatric subjects in training data [60], (ii) the application of distinct AASM scoring rules for children [8], and (iii) the absence of age as an input variable in both U-Sleep and YASA, which limits their ability to learn these differences. Although these issues are specific to sleep-scoring, similar challenges—where algorithms under-perform on under-represented data or due to omitted variables—are prevalent across domains, highlighting the usefulness of our approach. Performance was also systematically affected by AHI and PLMI, both of which indicate the severity of breathing disturbances and movement events during sleep. These factors led to lower accuracy and increased performance variability across both models. Additionally, U-Sleep exhibited significantly lower performance for male subjects, whereas none of the studied factors significantly influenced the inflation parameters. While this is a desirable outcome, it is likely due to the small number of extreme cases (<5 PSGs) of perfect-prediction or complete-misclassification in our dataset. The varying levels of sleep-scoring performance across different subpopulations highlight the need for human oversight, ensuring the validation of algorithmic predictions [95], and hence promoting equitable care across all age groups.

Further, we applied our approach to quantify the distribution of systematic error (*bias*) across algorithms concerning sleep-staging-derived clinical markers (e.g., sleep stage proportions, REM latency), which play a key role in diagnostics. Bias was quantified as the difference between values of markers based on sleep-stages predicted by the algorithms and those derived from human-scored hypnograms, accounting for the same sensitive attributes (age, gender, AHI, PLMI) as in the performance analysis. For both algorithms and all 13 evaluated markers, we identified significant non-linear spline effects of age on both bias expectation and its variability. This extends our findings on age-related performance variations to systematic distortions in derived clinical markers, indicating both systematic over- or underestimation of reference values based on subject age and changes in precision, reflected in the dispersion of expected bias quantiles. We presented in detail the ability of algorithms to accurately estimate wakefulness proportion (W%), which is functionally linked to several other PSG markers, including sleep latency, total sleep time, WASO, number of awakenings, sleep efficiency, and the proportion of remaining sleep stages. Across all age groups, U-Sleep underestimated W%, whereas YASA overestimated it, with both bias magnitude and variability increasing in older subjects and with AHI and PLMI. Even more pronounced

bias was observed in pediatric subjects, likely caused by the same challenges as identified for performance metrics. Bias in W% propagated to other PSG markers, where U-Sleep's underestimation of W% led to a corresponding overestimation of total sleep time and sleep efficiency, while the opposite trend was observed in YASA, which overestimated W%. Detailed results for each algorithm, clinical marker, and arbitrary values of sensitive attributes (age, gender, AHI, PLMI) can be further explored through our interactive app, enabling the assessment of algorithmic bias and performance across different subpopulations.

Interestingly, direction of biases, i.e., whether a clinical marker is over- or underestimated, often differed between U-Sleep and YASA. For instance, W% tends to be underestimated by U-Sleep and overestimated by YASA, with this effect propagating into downstream statistics such as TST, WASO, etc. These opposing trends suggest that the observed biases likely stem from differences in the algorithms themselves, e.g., due to imbalanced training data, architectural characteristics of U-Sleep, or selected input features in YASA, rather than from the BSWR dataset alone, which likely suffers from selection bias. While the primary aim of this study is to present and demonstrate a novel validation framework, robust conclusions about generalizable algorithmic bias would require prospective studies using representative populations, as mandated by regulatory authorities for the certification of AI-based medical software (cf. MDR, EU AI Act).

The results of performance and bias quantification confirmed a trend of increased error rates and greater variability at the extremes of age (i.e., pediatric and elderly individuals), as well as with elevated AHI and PLMI values. Clinical indices (AHI, PLMI) are indicative of poorer sleep health, which tends to correlate with age, and all that often presents with altered sleep architecture and signal artefacts [18], [78], [79], [139], [151], [152]. The increased variability can plausibly be attributed to the well-documented rise in inter-scorer disagreement when scoring PSGs from clinically complex subjects [67]–[69]. Rather than questioning whether the "truth" in sleep staging lies with the human scorer or the algorithm—where the latter is often benchmarked against soft-consensus scores from multiple independent experts and has, in some studies, even outperformed individual scorers [61], [70], [72], and the former represents the de facto standard in real-world clinical workflows, where scoring is typically performed by a single trained scorer in alignment with regulatory expectations under the MDR and EU AI Act—our work emphasizes the importance of quantifying these trends.

Finally, we evaluated the diagnostic utility of derived markers, despite their inherent biases. By training five simple classifiers (e.g., LDA, Random Forest) to distinguish OSA subjects from healthy controls, we found no statistically significant difference in predictive performance (e.g., AUROC) between classifiers trained on physician-scored markers and those trained on algorithm-predicted markers. Although we demonstrated that algorithm-derived markers exhibit systematic biases (e.g., U-Sleep underestimates W%, while YASA overestimates it), the classifiers adapted to these shifts due to their systematic nature, ultimately achieving comparable performance. This finding does not contradict the necessity of validation and bias quantification but rather highlights that when systematic errors are present in predicted markers, they can still retain the same predictive capability as human-scored markers, making them comparably useful for screening applications.

4.5 Conclusion

This study advances methods for identifying and quantifying algorithmic performance and bias, offering a framework that allows for the evaluation of external factors on error and bias, the modelling of non-linear effects, and the application of standard statistical tests to assess their significance. Applied to automatic sleep-scoring, our approach highlighted primarily age-related performance shifts and biases. Our results suggest that omitting subjects' age in sleep-scoring algorithms not only ignores biologically distinct EEG patterns [18], [78], [79] but also prevents learning different scoring rules applied to pediatric and adult cases- as defined by AASM guidelines [8]. Our evaluations confirm common sources of algorithmic bias, including under-representation of certain subgroups (e.g., children), omitted variables (e.g., age), and the use of observational rather than experimental data—the latter being costly but crucial for bias mitigation. Using our framework for identifying sub-populations affected by

pronounced algorithmic bias can help guide the selection of training or fine-tuning data for its iterative mitigation. In conclusion, our study is convinced of the benefit and usefulness of automatic sleep-scoring algorithms, both in clinical settings and the expanding consumer device market. To ensure fair and reliable predictions, it is essential to exercise caution in their use, adhere to emerging regulatory frameworks, incorporate human oversight, and raise awareness among physicians regarding their technical limitations and potential biases.

4.6 Limitations

Our study has several limitations. First, while our framework quantifies and tests for bias, it does not directly address mitigation strategies, which could involve post-processing techniques or adjustments in training procedures [161], [162]. Future extensions could benefit from integrating dedicated fairness toolkits such as Fairlearn [163] or AIF360 [164]. Second, our analysis considered only four sensitive attributes (age, gender, AHI, and PLMI) when quantifying performance and bias. Although these variables roughly cover an individual's health status, considering specific diagnoses and other clinical variables may help for more precise bias quantification. Since our framework is model-based, there is a risk that bias could persist within the bias-quantification model itself due to unaccounted confounders. While our method provides a structured approach, selecting the appropriate model remains a challenge, best addressed through a combination of domain expertise and high-quality data. Furthermore, the study dataset (BSWR) is observational and non-randomized, and while it includes scoring patterns from over 60 physicians, individual scorers evaluated different numbers of PSGs, potentially influencing the distribution of scoring tendencies in the dataset. Additionally, each PSG in the BSWR dataset was scored by a single expert, which prevents direct comparisons against consensus and limits our ability to analyze inter-scorer agreement or evaluate algorithm performance relative to scoring variability across multiple human annotators. While the use of a single out-of-domain dataset (BSWR) enabled an in-depth illustration of our proposed framework, validating the generalizability of the quantified biases would benefit from applying the approach to additional, ideally prospectively collected and representative datasets, and from re-training both algorithms (U-Sleep, YASA) on shared cohorts. We aim to focus our future work on addressing the study limitations.

Chapter 5

Novel Digital Markers of Sleep Dynamics: Causal Inference Approach Revealing Age and Gender Phenotypes in Obstructive Sleep Apnea

Abstract

Despite evidence that sleep-disorders alter sleep-stage dynamics, only a limited amount of these parameters are included and interpreted in clinical practice, mainly due to unintuitive methodologies or lacking normative values. Leveraging the matrix of sleep-stage transition proportions, we propose (i) a general framework to quantify sleep-dynamics, (ii) several novel markers of their alterations, and (iii) demonstrate our approach using Obstructive Sleep Apnea (OSA), one of the most prevalent sleep-disorder and a significant risk factor. Using causal inference techniques, we address confounding in an observational clinical database and estimate markers personalized by age, gender, and OSA-severity. Importantly, our approach adjusts for five categories of sleep-wake-related comorbidities, a factor overlooked in existing research but present in 48.6% of OSA-subjects in our high-quality dataset. Key markers, such as NREM-REM-oscillations and sleep-stage-specific fragmentations, were increased across all OSA-severities and demographic groups. Additionally, we identified distinct gender-phenotypes, suggesting that females may be more vulnerable to awakenings and REM-sleep-disruptions. External validation of the transition markers on the SHHS database confirmed their robustness in detecting sleep-disordered-breathing (average AUROC = 66.4%). With advancements in automated sleep-scoring and wearable devices, our approach holds promise for developing low-cost screening tools for sleep-, neurodegenerative-, and psychiatric-disorders exhibiting altered sleep patterns.

Keywords:
Sleep Dynamics, Digital Markers, Obstructive Sleep Apnea, Dirichlet Regression, Causal Inference, Sleep Disorders, Polysomnography

5.1 Introduction

The clinical sleep study (polysomnography, PSG) involves comprehensive overnight monitoring of body biosignals, including electroencephalogram (EEG), electrocardiogram (ECG), electromyogram (EMG), and others. Medical personnel evaluate the PSG following guidelines of the American Academy of Sleep Medicine (AASM) [8], focusing on the detection of complete and partial breathing arrests (i.e., apneas and hypopneas), movement events, and notably, categorizing stages of sleep. Sleep scoring - conventionally done manually for each 30-second window (epoch) of the biosignals recorded - differentiates between five sleep-wake stages: wakefulness (W), rapid-eye-movement (REM) sleep, and three other non-REM

(N1, N2, N3) sleep-states. Such a structured sleep-scoring (hypnogram) forms a basis for the PSG report, providing information on basic markers (e.g., sleep efficiency, % of sleep-stages, REM latency) that relate to sleep quality and may also indicate certain sleep disorders [18], [139], [165].

Sleep and its markers have a complex relationship with individuals' age and may vary by gender [166]. Several meta-analyses have made considerable efforts to establish normative values of sleep markers in healthy individuals [9], [17]. However, the validity of certain estimates might be questionable due to inappropriate statistical evaluations of the individual studies whose results were pooled [167]. For instance, REM latency, as a time-to-event phenomenon subject to censoring, is best quantified using survival techniques rather than mean comparisons. Similarly, the % of sleep-stages, which are interdependent, should be assessed by compositional methods. Proper techniques enabling unbiased estimation are however rarely applied. Quantification of normative ranges and changes in sleep markers in diseased subjects is even more challenging. The observational study design of PSG databases, typically including non-randomized symptomatic subjects, introduces a high degree of confounding [168]. This results in an imbalanced prevalence of individuals with different clinical statuses and distributional shifts in their demographic characteristics. These factors make it difficult to separate the effects of natural ageing from the effects of particular disorders on sleep parameters. The unaddressed confounding, difficulty in assessing data of patients who often suffer from several sleep disorders simultaneously, and the use of not always appropriate statistical approaches are major challenges that increase the risk of biased conclusions even in the analysis of well-established PSG markers.

While differences in sleep-stage dynamics are evident for certain sleep disorders, such as increased sleep fragmentation in Obstructive Sleep Apnea (OSA) [25], [169], or a short REM latency in narcoleptic patients [170], the clinical PSG report has, so far, included only a limited number of dynamics-related markers. This includes sleep and REM latencies and the absolute counts of sleep-stage transitions or awakenings [8]. While latencies target the first (tens of) minutes of the night, the overall numbers of transitions/awakenings are proportional to sleep duration and may not sufficiently capture more complex patterns of sleep dynamics that may be specific to individual sleep disorders. Although counts of transitions and awakenings are sometimes normalized as indices per hour, stage-specific dynamics—such as REM-related continuity and transitions—are typically overlooked, despite their potential to reveal disorder-specific patterns of sleep disruption. Their limited incorporation into clinical PSG reports is largely due to the absence of standardized methodologies, normative values, and intuitive frameworks to support clinical interpretation. Recognizing these limitations, significant research has been conducted to comprehensively explore sleep-stage dynamics in various modalities. These studies, which date back to the 1980s, exhibit heterogeneity in terms of subject demographics, clinical diagnoses, and the methodologies employed [171]. Two main investigative directions have emerged: (i) *focusing on the transitions between sleep stages*, and (ii) *focusing on the duration of sleep stages*. The perspectives of these two seemingly distinct but strongly interrelated areas are discussed in the following two separate paragraphs, highlighting the contribution of the most impactful studies.

Research on sleep-stage transitions has evolved rapidly, beginning with one of the earliest mathematical models by Kemp (1986), who quantified transition intensities in 23 healthy males aged 18-30 [98]. Yassouridis (1999) followed by exploring the relationship between transition intensities and plasma cortisol levels in 30 males aged 20-30 [99]. Several studies identified associations between transition rates and clinical symptoms. For instance, Burns (2008) observed increased sleep fragmentation and transitions into N3 in 15 females with fibromyalgia syndrome (mean ± standard deviation (SD) age of 42.5 ± 12.9), contrasting with age- and gender-matched controls [100]. Laffan (2010) found a significant association between transition rates and self-reported sleep quality in a large cohort from the Sleep Heart Health Study (SHHS) database, consisting of 5684 participants (47.2% males, all aged over 40) [101]. The existing research extends to specific conditions such as chronic fatigue syndrome, where Kishi (2008) reported abnormal REM transitions in 22 female patients (aged 42 ± 8) in comparison to healthy controls of similar demographics [102]. Further exploring clinical implications, Kim (2009) found differences in sleep-stage dynamics between nights with and without CPAP therapy in 113 OSA subjects (aged 54.0 ± 11.7, 16 females) [103]. Wei (2017) documented increased N2-to-W/N1 transitions in 46 insomnia patients (aged

50.3 ± 13.6, 8 males) compared to age- and gender-matched controls, indicating altered sleep patterns [104]. In addition, Schlemmer (2015) analyzed first- and second-order sleep-stage transitions across 4 groups of subjects (young vs old, healthy vs disorder), highlighting the varied impacts of ageing and pathological conditions [105]. Yet, the disordered subjects represented a pool of various sleep and psychological conditions, and the findings cannot be attributed to a specific diagnosis. Recently, Wachter (2020) utilized MANOVA adjusted for age, gender, and BMI, to evaluate differences in the 25 most common second-order transitions in different severities of OSA compared to healthy subjects, demonstrating associations with demographic and clinical factors [106]. The significant findings primarily related to wake and light-sleep (N1, N2) oscillations, when comparing severe-OSA and healthy. An innovative yet not diagnosis-oriented approach by Yetton (2018) applied a Bayesian network to model transitions as well as stage durations in 3,202, according to exclusion criteria, healthy subjects (mean age of 62.5, 60% males). The prediction-oriented results demonstrated the highest accuracy (62.3%) in the identification of the current stage based on the previous 2 stages, the duration of the last stage, and no consideration of age, gender, or BMI [107].

Another perspective in understanding sleep dynamics focuses on the quantification of sleep stage durations, providing insights into the temporal characteristics of individual sleep-wake periods. Lo (2002) initiated this research direction by examining sleep-wake dynamics in 20 healthy subjects (aged 23-57, 9 males), revealing different characteristics between sleep and wake periods' duration and advocating for their modelling using power law distributions [108]. Building on this, Penzel (2003) applied power-law models to quantify sleep-stage durations in both healthy and disordered subjects, identifying reduced duration and hence more fragmented sleep in sleep-apnea subjects [109] (with no specific demographic details provided). Following that, Norman (2006) exploited survival techniques and revealed decreased sleep continuity when comparing 10 mild and 10 moderate/severe subjects with sleep-disordered-breathing (SDB) against 10 normal subjects [110]. The analysis did not consider subjects' age, which was significantly higher in disordered subjects. Chervin (2009) compared sleep architecture in 48 children (aged 5-12.9) with sleep-disordered breathing to healthy controls, finding a significant decrease in the duration of N2 and REM [111]. Bianchi (2010) employed multi-exponential fitting to analyze sleep-stage durations across 376 predefined controls (aged 68.2 ± 6.3, 35.6% males), in comparison to 496 mild-OSA (aged 63.8 ± 0.3, 60% males), and 338 severe-OSA (aged 63.7 ± 10.5, 70.7% males) subjects from the SHHS database [112]. They report accelerated decay rates in W, NREM, and REM among OSA subjects, suggesting a larger sleep fragmentation and shorter stage bouts. Notably, despite considerable age and gender differences within its sample (35.6% vs 70.7% males in healthy vs severe-OSA), the study did not adjust for them. Klerman (2013) investigated durations of sleep-wake states in healthy subjects and identified an age-related decline of NREM-sleep continuity [113]. A comparison of sleep-stage duration by Kishi (2020) in sleep bruxism (SB) patients (aged 23.3 ± 1.1, 6 males) and matched controls showed that despite no differences in the prevalence of sleep-stages (except for N1), the SB subjects differed in several parameters describing their dynamics, particularly related to an increased REM fragmentation and hence reduced duration of REM-bouts [114].

By analysing sleep-stage transitions [98]–[107] or by characterizing their duration [108]–[114], all of these studies highlight the importance and clinical utility of analysing sleep dynamics across a wide range of disorders. Although most of the studies focus on one of these two aspects, it is important to point out that their nature is functionally linked as the lower transition probability relates to an increased bout duration [172], [173]. The existing research works have variously addressed the complexities of confounding and the selection of appropriate statistical models. The majority of studies concurred on the need to control for age and gender or limit the demographic ranges to ensure a homogeneous group of study participants. In existing studies, this is achieved by using stratified analysis with (M)ANOVA (e.g., [105], [106], [112]), regression adjustment (e.g., [101]), or selecting matched individuals (e.g., [100], [104], [114]). The simplicity of the first two approaches, typically comparing the effect of exposure (such as OSA) on the outcome (e.g., sleep dynamics) against unexposed healthy controls, is offset by its susceptibility to confounding bias [174]. Analyzing non-randomized observational PSG databases, which typically include older, symptomatic individuals, complicates the separation of confounder effects (of age, gender) from the exposure (disorder). In contrast, while the matching approach helps a lot to reduce the bias [175],

it is generally applied within smaller subject cohorts. This limitation arises from the challenges of finding individuals with matched characteristics within typically imbalanced clinical databases of limited size.

Our study introduces a comprehensive framework for quantifying sleep dynamics, demonstrated on OSA but applicable to other (sleep) disorders. OSA, one of the most prevalent sleep disorders and a significant risk factor affecting up to 17% of the general adult population [149], serves as a use-case to showcase the framework's versatility. Building on existing research and addressing its limitations, our framework—depicted in Figure 5.1 and detailed in *Methods*—fulfils several key objectives:

- **Data acquisition**, Figure 5.1a: Leveraging a high-quality, heterogeneous observational clinical database, we identified OSA and healthy subjects (aged 6-91 years) based on the clinical gold-standard of *conclusive* diagnosis. Consistent with the literature (e.g., [101], [105], [106], [112], [149]), we identified age and gender as the primary confounders. The subjects' sleep was summarized through AASM-scored hypnograms, forming the basis for proposing and deriving novel digital markers of sleep and its dynamics. The information about sleep comorbidities was also considered to adjust our framework for additional possible confounders. The importance of the need for comorbidity-adjustment can be underscored by the fact that 48.6% of OSA subjects in our dataset had at least one sleep-wake comorbidity among their conclusive diagnoses.

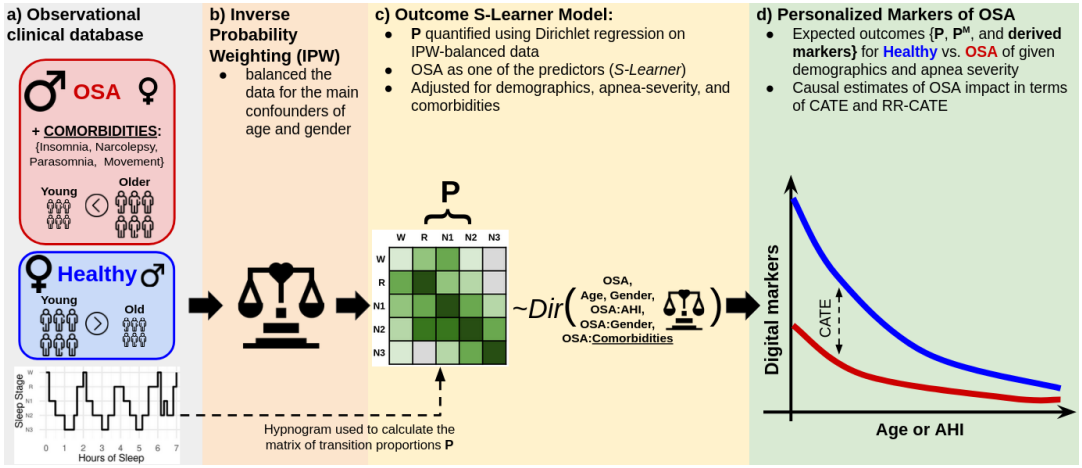
- **Balancing confounders**, Figure 5.1b: To address confounding of age and gender, exhibiting distributional overlap between OSA and healthy subjects, we applied *Inverse Probability Weighting (IPW)* (c.f., [176]–[178]) that ensured balanced comparisons between the OSA and healthy groups, regarding the main confounding factors. In short, IPW aims to mathematically re-weight the original dataset, as it was matched regarding the confounders considered (such as age).

- **Sleep dynamics modeling**, Figure 5.1c: Utilizing hypnograms, we propose a novel *"sleep fingerprint"*, a matrix **P** of sleep-stage transition proportions. As first ones in the field, respecting the interdependencies between individual dimensions of transition proportions **P** (that sum-up to 100%), we quantified them jointly using Dirichlet regression [179], a method well-suited for the compositional nature of **P**, within a causal S-Learner framework [180] applied to IPW-balanced data. The idea of causal S-Learner is to extrapolate outcomes for *"conditioned"* (OSA) vs control (healthy) subjects for arbitrary values of predictors. This approach enables the estimation of changes in sleep (dynamics) across different ages, OSA-severities (AHI), and the previously understudied interplay of OSA with gender and sleep-wake-related comorbidities.

- **Digital marker quantification**, Figure 5.1d: Finally, by exploiting the estimated model (1.c), we quantify not only the estimated effects of OSA on **P** but also derive several novel digital markers. These markers capture the disorder's impact on sleep, sleep-stage dynamics and also durations, personalized for arbitrary values of predictors (such as age, gender, apnea-severity), and are presented in terms of *Conditional Average Treatment Effect* (CATE) and *Risk-Ratio CATE* (RR-CATE) [181], standing for absolute and relative comparisons of expected outcomes (such as specific stage-transitions) for OSA and healthy, respectively.

Our framework integrates the two main branches of sleep dynamics research—quantification of sleep-stage transitions and durations—by demonstrating their interconnectedness and enabling their simultaneous quantification. Our study is the first in the field to rigorously account for the interactions between OSA, gender, and a wide range of comorbidities, providing a deeper understanding and less biased estimates of how OSA impacts sleep across various ages, genders, and apnea-severity levels. As demonstrated in our results, the quantified effects and markers of OSA can be leveraged to: (i) *explain*—by establishing normative values for sleep parameters tailored to different demographic profiles and OSA severity; and also (ii) *predict*—by training models capable of identifying OSA subjects based solely on observed demographics and sleep-stage dynamics. The results are publicly accessible through an interactive online app, fostering a broader scientific exploration and discussion.

Figure 5.1: Graphical overview of the implemented approach for quantifying sleep-stage dynamics.



Notes: Part a): The study utilized observational data, including hypnograms of subjects with a conclusive diagnosis of either Obstructive Sleep Apnea (OSA) or healthy status. The illustration highlights differences in the overall prevalence of OSA (OSA-affected > healthy) concerning gender (male predominance in OSA), age (higher OSA prevalence in older subjects), and comorbidities (not present in healthy subjects). Part b): Inverse Probability Weighting (IPW) is applied to balance the data for the primary confounders of age and gender, having distributional overlap between OSA and healthy subjects. Part c: A sleep fingerprint matrix **P** of sleep-stage transition proportions is modelled using Dirichlet regression within a causal S-Learner framework to capture the effects of OSA, its severity (Apnea-Hypopnea Index, AHI), age, gender, and comorbidities. Part d): The framework quantifies digital markers of OSA (raw **P**, $\mathbf{P}^M$ as the normalized Markovian **P**, and derived quantities such as sleep fragmentation), personalized for subjects' demographics, OSA severity, and comorbidities, and presented in terms of Conditional Average Treatment Effect (CATE) and Risk-Ratio CATE (RR-CATE).

5.2 Materials and Methods

This section details the study data, introduces the matrix of sleep-stage transition proportions as a foundational digital marker, and explores its properties alongside several novel sleep markers. Additionally, we outline the technical framework, which leverages causal inference tools to minimize bias in the conclusions of this observational study, and present a use case examining the effects of OSA.

5.2.1 Data

Berner Sleep Data Base (BSDB)

For the primary evaluations (such as estimating the effects) of our study, we exploited the clinical *Berner Sleep Data Base* (BSDB) from Inselspital, University Hospital Bern. We considered a subset of 62 healthy subjects (aged 0-71 years) with excluded existing clinical conditions undergoing PSG as controls in several historical studies, and a total of 560 individuals having OSA (aged 2-81 years, including 2 pediatric cases aged < 18 years) as one of their conclusive diagnoses, made by physicians considering all test-based diagnoses (e.g., actigraphy- or PSG-based), clinical anamnesis, and the context. The PSG signals were recorded at 200 Hz and scored manually according to the AASM rules [8]. To align older recordings scored by Rechtschaffen and Kales [75] rules with the AASM standard, N3 and N4 stages were merged into N3. To prevent bias due to possibly longer sleep-onset in the unfamiliar clinical setting, a part of the PSG recording and hypnogram before the first sleep was cut off. Further, recordings with total sleep time <180 minutes, >5% of the time with lights-on, no sleep-stage transitions, and subjects with breath control or ventilation therapy introduced, or undergoing split-night PSG evaluations were excluded. For the basic statistical description of BSDB in Table 5.1, we considered 3 groups of OSA subjects: mild (O1) with AHI $\in [5, 15)$, moderate (O2) with AHI $\in [15, 30)$, and severe (O3) with AHI ≥ 30.

Table 5.1: Comparison of demographics, sleep metrics, and prevalence of sleep comorbidities among healthy and (mild, moderate, severe) OSA subjects in the BSDB dataset.

	H: Healthy	O1: Mild OSA	O2: Moderate OSA	O3: Severe OSA	Significant Pairs
DEMOGRAPHICS:					
Subjects [count]	62	238	164	158	
*Males	25 (40.3)	166 (69.7)	117 (71.3)	127 (80.4)	O1H, O2H, O3H
*Females	37 (59.7)	72 (30.3)	47 (28.7)	31 (19.6)	O1H, O2H, O3H
†Age	34.9 (18)	50.6 (14.9)	53.8 (14.7)	58 (11.9)	O1H, O2H, O3H, O3O1
SLEEP METRICS:					
†TST [minutes]	370.3 (62.9)	345.5 (74.3)	344.1 (76.4)	321.4 (64.2)	O2H, O3H, O3O1, O3O2
†Efficiency [%]	88.4 (6.6)	83.4 (11.7)	81.3 (12.2)	78.9 (12.5)	O2H, O3H, O3O1
†Sleep latency [minutes]	16.5 (15.4)	12.9 (19)	16.6 (24)	15.7 (22)	O1H
†REM latency [minutes]	113.4 (50.7)	138.9 (80.1)	124.4 (72.6)	148.6 (86.3)	-
†Hourly transitions [$\frac{N}{hour}$]	16.2 (4.9)	20.6 (5.6)	22.2 (6.1)	25.9 (7.9)	O1H, O2H, O3H, O2O1, O3O1, O3O2
†Hourly awakenings [$\frac{N}{hour}$]	2.4 (1.2)	3.2 (1.7)	3.3 (1.7)	4 (2.8)	O1H, O2H, O3H, O3O1
†W [%]	11.6 (6.6)	16.6 (11.7)	18.7 (12.2)	21.1 (12.5)	O1H, O2H, O3H, O3O1
†N1 [%]	9.4 (6.3)	13.6 (7.6)	16.3 (9.1)	24.1 (13.3)	O1H, O2H, O3H, O2O1, O3O1, O3O2
†N2 [%]	40.6 (10.3)	39.9 (10.3)	36.1 (10.7)	34.1 (13.2)	O2H, O3H, O2O1, O3O1
†N3 [%]	21.7 (8.7)	16.9 (9.7)	16.4 (11.1)	10.2 (7.9)	O1H, O2H, O3H, O3O1, O3O2
†REM [%]	16.7 (7.1)	13.1 (6.7)	12.5 (6.4)	10.4 (6.5)	O1H, O2H, O3H, O3O1, O3O2
SLEEP COMORBIDITIES:					
*No comorbidity	62 (100)	94 (39.5)	85 (51.8)	109 (69)	O3O1, O3O2
*Single comorbidity	0 (0)	52 (21.8)	37 (22.6)	30 (19)	-
*Multiple comorbidities	0 (0)	92 (38.7)	42 (25.6)	19 (12)	O2O1, O3O1, O3O2
*Insomnias	0 (0)	44 (18.5)	22 (13.4)	15 (9.5)	-
*Narcolepsy type 1	0 (0)	11 (4.6)	7 (4.3)	4 (2.5)	-
*Other hypersomnias	0 (0)	88 (37)	37 (22.6)	15 (9.5)	O2O1, O3O1, O3O2
*Parasomnias	0 (0)	28 (11.8)	25 (15.2)	22 (13.9)	-
*Movement-related	0 (0)	23 (9.7)	13 (7.9)	12 (7.6)	-
*Circadian-rhythm-related	0 (0)	2 (0.8)	3 (1.8)	1 (0.6)	-

Notes: Variables denoted with * are binary, summarized as count (percentage), N (%), and significantly different pairs are listed, following a significant chi-squared independence test and pairwise posthoc proportions test. Healthy subjects were excluded from the comorbidities comparisons as they had no comorbidities. Variables denoted with † are continuous, summarized as mean (standard deviation), $\mu(\sigma)$, and significant pairs are listed following a significant Kruskal-Wallis test and pairwise Wilcoxon posthoc test. All posthoc pairwise comparisons were performed with Bonferroni corrections at the significance level of 0.05.

Most sleep metrics and demographics differ significantly between healthy individuals and OSA groups, as well as across different OSA severity levels. There is a clear trend of increasing age and % of males from healthy to more severe OSA, which is also associated with changes in sleep architecture, such as decreased sleep efficiency and reduced N3 and REM %. Separating the effects of these demographic shifts from the effects of OSA is a key challenge, addressed using a causal inference below.

Ethics Approval and Consent The secondary usage of the dataset was approved by the local ethics committee (Kantonale Ethikkommission Bern [KEK]-Nr. 2022-00415), ensuring compliance with the Human Research Act (HRA) and Ordinance on Human Research with the Exception of Clinical Trials (HRO). All methods were carried out in accordance with relevant guidelines and regulations. Written informed consent was obtained from all participants, as part of the general consent process introduced at Inselspital in 2015. Data were maintained with confidentiality throughout the study.

Sleep Heart Health Study (SHHS)

The Sleep Heart Health Study (SHHS) is a large, multi-centre cohort study designed to investigate the relationship between sleep-disordered breathing and cardiovascular outcomes [182], [183]. SHHS1 includes baseline polysomnography (PSG) data collected from 5,804 unique subjects aged 39–90 years, while SHHS2 provides follow-up PSG data for 2,651 subjects aged 44–90 years. Following the same criteria as in BSDB, we included only subjects with total sleep time (TST) > 180 minutes. After this selection, SHHS1 retained 5,734 subjects (mean age 63.1 ± 11.2 years, 47.6% male), and SHHS2 included 2,621 subjects (mean age 67.5 ± 10.3 years, 46.1% male).

SHHS1 and SHHS2 were utilized to independently evaluate the predictive power of individual sleep-stage transition proportions, forming the foundation for deriving novel sleep markers in identifying subjects with sleep-disordered breathing. These analyses provide robust external validation of the effectiveness of these transition proportions in the predictive task, which underscores their clinical relevance.

For both BSDB and SHHS datasets, the definition of the Apnea-Hypopnea Index (AHI) used aligns with the National Sleep Research Resource (NSRR) harmonization [183]: AHI = (All apneas + hypopneas with $\geq$30% nasal cannula [or alternative sensor] reduction and $\geq$3% oxygen desaturation or with arousal) per hour of sleep, which follows clinical guidelines [8].

5.2.2 Matrix P of sleep-stage transition proportions: a basic sleep marker

Our framework proposes the use of a flexible digital marker—a sleep fingerprint—that, based on the observed sleep stages of a subject, enables the derivation of both established and novel PSG parameters, quantifying various sleep characteristics that may be specific to different sleep conditions. The basis for achieving this is the hypnogram, which represents the sequence of sleep-wake stages (W, N1, N2, N3, REM) throughout the night. While sleep dynamics in clinical PSG reports are currently limited to the total counts of transitions and awakenings, this can be easily extended by the 5 x 5 matrix of sleep-stage transition proportions **P**. Let us denote the total number of epochs in the patient's hypnogram (starting from sleep-onset) as N^E, and the number of transitions from stage i to j as N^{ij}. Each cell p_{ij} of **P** can then be expressed as:

$$p_{ij} = \frac{N^{ij}}{N^E} = P(\text{next stage} = j, \text{current stage} = i) = P(i \rightarrow j) \qquad \forall i, j \in \{\text{W, N1, N2, N3, REM}\}, \tag{5.1}$$

indicating the empirical probability (proportion, %) of observing a transition from stage i to j ($i \rightarrow j$), relative to all the transitions observed in the hypnogram. In the following, we highlight three main dimensions of the clinical relevance of **P**.

P recovers the majority of clinically established PSG markers

For example, summing up the column transition proportions of **P** yields the overall percentage of sleep stages:

$$\text{stage } j\,\% = p_{*,j} = \sum_{i \in \{\text{W, N1, N2, N3, REM}\}} p_{i,j} \quad \forall j \in \{\text{W, N1, N2, N3, REM}\}. \tag{5.2}$$

In addition, other clinically commonly used PSG markers can be easily derived by considering relevant proportions and the *Total Sleep Time* (TST), $\text{TST} = \frac{N^E}{2}$, in minutes. For example, *Sleep Efficiency* (SE), quantifying the percentage of sleep after its onset, can be calculated as $\text{SE} = \sum_{j \in \{\text{N1, N2, N3, REM}\}} p_{*,j} = 1 - p_{*,\text{W}}$. The *Wake After Sleep Onset* (WASO) minutes can be computed as $\text{WASO} = \frac{N^E}{2} p_{*,\text{W}}$. The *Number of Awakenings* (NoA) can be determined by $\text{NoA} = N^E \sum_{i \in \{\text{N1, N2, N3, REM}\}} p_{i,\text{W}}$. Finally, the *Number of Transitions* (NoT) is given by $\text{NoT} = N^E \sum_{i \in \{\text{W, N1, N2, N3, REM}\}} (1 - p_{i,i})$.

P allows derivation of novel PSG markers

The aggregation of **P**-dimensions offers great flexibility to derive several novel and highly intuitive digital markers of sleep and its dynamics. Considering a set of sleep-states, $\mathcal{S} = \{\text{N1, N2, N3, REM}\}$, we propose and in results also evaluate the following.

Total Awakenings, the probability of transitioning from any sleep-state ($\mathcal{S}$) to wakefulness:

$$P(\mathcal{S} \to \text{W}) = \sum_{i \in \mathcal{S}} p_{i,\text{W}} = p_{\text{N1,W}} + p_{\text{N2,W}} + p_{\text{N3,W}} + p_{\text{REM,W}}, \tag{5.3}$$

Light-sleep Awakenings, the probability of transitioning from light sleep (N1, N2) to wakefulness:

$$P(\text{Light-sleep} \to \text{W}) = p_{\text{N1,W}} + p_{\text{N2,W}}, \tag{5.4}$$

Deep-sleep Awakenings, the probability of transitioning from deep sleep (N3) to wakefulness:

$$P(\text{N3} \to \text{W}) = p_{\text{N3,W}}, \tag{5.5}$$

REM Awakenings, the probability of transitioning from REM sleep to wakefulness:

$$P(\text{REM} \to \text{W}) = p_{\text{REM,W}}, \tag{5.6}$$

NREM-REM Oscillations, sum of probabilities for transitions between NREM sleep stages and REM sleep:

$$P(\text{NREM} \rightleftarrows \text{REM}) = \sum_{(i,j) \in \{N1,N2,N3\} \times \{REM\}} p_{i,j} \tag{5.7}$$

Light-sleep Oscillations, sum of probabilities for transitions between the light sleep stages (N1, N2):

$$P(\text{N1} \rightleftarrows \text{N2}) = p_{N1,N2} + p_{N2,N1}, \tag{5.8}$$

Sleep Compactness, the total probability of staying within any (non-wake) sleep stages:

$$P(\text{Sleep Compactness}) = \sum_{(i,j) \in \mathcal{S} \times \mathcal{S}} p_{i,j}, \tag{5.9}$$

Sleep Fragmentation, the total probability of switching between wakefulness and sleep states:

$$P(\text{Sleep Fragmentation}) = \sum_{i \in \mathcal{S}} (p_{W,i} + p_{i,\text{W}}), \tag{5.10}$$

Sleep-stage Compactness, the sum of probabilities of staying within the same (non-wake) sleep stages:

$$P(\text{Sleep-stage Compactness}) = \sum_{i \in \mathcal{S}} p_{i,i}, \tag{5.11}$$

Sleep-stage Fragmentation, the probability of transitioning from one (non-wake) sleep stage to a different one:

$$P(\text{Sleep-stage Fragmentation}) = \sum_{\substack{(i,j)\in\mathcal{S}\times\mathcal{S} \\ i\neq j}} p_{i,j} \tag{5.12}$$

Stage-specific Compactness and Fragmentation, for each sleep stage i, the probability of staying in the same stage and the probability of switching to any other sleep stage, respectivelly:

$$P(i\text{-th stage Compactness}) = p_{i,i}, \quad P(i\text{-th stage Fragmentation}) = \sum_{j:i\neq j} p_{i,j} \quad \forall i \in \{\text{W, N1, N2, N3, REM}\} \tag{5.13}$$

Each metric from Eq. 5.3-5.13 expands the standard clinical PSG markers and focuses on a specific sleep pattern. Their quantification requires no additional effort once the subject has undergone the PSG study and the hypnogram is available.

P bridges stage-transitions and durations-oriented sleep dynamics research.

Normalizing **P** so that each row sums to 1 (100%) yields a standard transition matrix, often utilized in Markovian models. We denote this matrix as $\mathbf{P}^M$, where M indicates it is Markovian. Each cell, $p^M_{i,j}$, corresponds to the conditional probability of transitioning to stage j after being in stage i:

$$p^M_{i,j} = P(\text{next stage} = j \mid \text{current stage} = i) = \frac{p_{i,j}}{p_{i,*}} = \frac{p_{i,j}}{\sum_{j\in\{\text{W, N1, N2, N3, REM}\}} p_{i,j}} \quad \forall i,j \in \{\text{W, N1, N2, N3, REM}\} \tag{5.14}$$

The key difference is that while **P** provides an overall view of the plausibility of individual transitions, $\mathbf{P}^M$ operates under the assumption that a given state has occurred and problematically evaluates the chances of (not-)switching the sleep-stage in the next epoch. Both **P** and $\mathbf{P}^M$ are interconnected and offering two perspectives on sleep-stage dynamics. Notably, the diagonal elements of $\mathbf{P}^M$ enable straightforward quantification of the sleep-stage durations, as they are exponentially distributed, $\mathcal{E}(\lambda) = \mathcal{E}(1 - p^M_{i,i})$, with the expected duration (ED) of each stage (over entire night):

$$\text{ED}_i = E(\text{duration of stage } i) = \frac{1}{\lambda} = \frac{1}{1 - p^M_{i,i}} \quad \forall i \in \{\text{W, N1, N2, N3, REM}\}, \tag{5.15}$$

known as the mean sojourn time. Due to the scoring of sleep in 30-second windows, these durations are measured in epochs.

5.2.3 Causal framework to quantify sleep-stage transition matrix P and effects of a disorder

The preceding sections have highlighted the utility of investigating the matrix **P** as a sleep-fingerprint, showing its relation to several clinically established PSG markers and its connection between stage-transition and stage-duration sleep dynamics research. Moreover, we introduced several novel markers derived from **P**. To quantify **P** and the derived markers, the next sections will present an approach that combines Dirichlet regression, well-suited for the compositional data of **P**, with elements of causal inference to address confounding. The key challenge in modeling **P** lies in respecting the compositional nature of the data, where the total of all percentages must sum to 100%. Ignoring this constraint, such as analyzing particular proportions separately with ANOVA, can lead to significant bias and counterintuitive outcomes. This issue is evident in some meta-analyses where, for example, aggregated percentages of sleep stages do not sum to 100%, as seen in Table 2 of [9]. This challenge must be addressed when modeling the proportions of sleep-stage transitions in **P**, which involve 25 compositional dimensions. Ensuring the outcomes are intuitive and correct is crucial for enabling their interpretation by medical professionals.

Dirichlet regression: model formulation and properties

The Dirichlet distribution is well-suited for modeling compositional data, such as percentages or the elements of **P**. For a random variable $Y = (Y_1, Y_2, ..., Y_D)$ representing proportions over D dimensions, the probability density function of the Dirichlet distribution is parameterized by a vector of positive reals $\alpha = (\alpha_1, ..., \alpha_D)$ and given by:

$$Dir(Y; \alpha) = \frac{1}{B(\alpha)} \prod_{d=1}^{D} Y_d^{\alpha_d - 1}, \tag{5.16}$$

where $B(\alpha)$ is the multivariate beta function ensuring normalization [179]. In Dirichlet regression, the logarithms of α are modeled as functions of covariates, adapting the distribution's characteristics based on predictor values:

$$\log(\alpha_d) = \beta_{d0} + \beta_{d1} X_1 + \cdots + \beta_{dK} X_K, \tag{5.17}$$

where $X = (X_1, ..., X_K)$ is a set of K covariates and $\beta_d = (\beta_{d0}, ..., \beta_{dK})$ a vector of regression coefficients for the d-th dimension. The expectation of each component Y_d, $E[Y_d]$, and the marginal effect of X_j on $E[Y_d]$, $\frac{\partial E[Y_d]}{\partial X_k}$, are directly influenced by all elements of X and α, reflecting the interdependencies of compositional data:

$$\begin{aligned} E[Y_d] &= \frac{\alpha_d}{\sum_{j=1}^{D} \alpha_j} = \frac{\exp(\beta_{d0} + \beta_{d1} X_1 + \cdots + \beta_{dk} X_K)}{\sum_{j=1}^{D} \exp(\beta_{j0} + \beta_{j1} X_1 + \cdots + \beta_{jk} X_K)}, \\ \frac{\partial E[Y_d]}{\partial X_k} &= E[Y_d] \left(\beta_{dk} - \sum_{j=1}^{D} \beta_{jk} E[Y_j] \right). \end{aligned} \tag{5.18}$$

A convenient property of the Dirichlet distribution is its ability to aggregate over several dimensions, allowing flexible quantification of measures based on the elements' summation. For example, aggregating dimensions i and j yields:

$$Y' = (Y_1, ..., Y_i + Y_j, ..., Y_D) \sim Dir(Y'; (\alpha_1, ..., \alpha_i + \alpha_j, ..., \alpha_D)). \tag{5.19}$$

Thus, Dirichlet regression is suitable for modelling **P**, and its aggregation property facilitates straightforward quantification of all markers derived from it (c.f., Eq. 5.2-5.13).

Causal elements

In contrast to randomized experiments, the analysis of observational data, such as those from PSG databases, is susceptible to confounding, due to varying distributions of characteristics (e.g., age), between treated/exposed/conditioned and healthy-control subjects. Our study, which aims to quantify changes in sleep parameters resulting from a sleep disorder, adopts the principles and standard notation of causal inference [181]. We define the *treatment/exposure/condition* variable T as an indicator of whether a subject suffers from a particular disorder of interest ($T = 1$), or is a healthy control ($T = 0$). **In line with the language of causal inference, the treatment within our study corresponds to the presence of OSA**. The *outcome* (Y) represents the sleep parameter investigated, such as **P**, while subject characteristics and potential confounders are denoted as X.

Potential outcomes framework and causal estimands. The potential outcomes framework asserts to each individual two hypothetical outcomes: $Y(1)$, under $T = 1$, and $Y(0)$, without exposure, $T = 0$. The *Individual Treatment Effect* (ITE), τ_i, is the difference between these outcomes, evaluating the causal effect of exposure (e.g., OSA) on subject's outcome (e.g., sleep):

$$\text{ITE} = \tau_i = Y_i(1) - Y_i(0). \tag{5.20}$$

The *Average Treatment Effect* (ATE) is the expected ITE, assessing the effect of T across the entire population:

$$\text{ATE} = E[\tau] = E[Y(1) - Y(0)]. \tag{5.21}$$

The *Conditional Average Treatment Effect (CATE)* assesses $\tau(x)$, standing for the treatment effect within a specific subgroup of the population characterized by covariates X, making it suitable to quantify personalized markers for different conditions:

$$\text{CATE}(X) = E[\tau(x)] = E[Y(1) - Y(0) \mid X], \tag{5.22}$$

The *fundamental problem of causal inference* is that only one of the two potential outcomes is observed for each individual, according to their treatment/exposure assignment T_i:

$$Y_i^{obs} = Y_i(T_i) = T_i Y(1) + (1 - T_i)Y(0), \tag{5.23}$$

making it impossible to directly calculate all hypothetical estimands (ITE/ATE/CATE) from observed data (Y_i^{obs}, T_i, X_i).

Personalized markers using CATE estimates. To estimate (C)ATE from observational data, advanced techniques are required to adjust for confounders and mimic a randomized experiment setting. One method exploits *Propensity Scores* (PS):

$$\pi(X_i) = P(T = 1|X_i), \tag{5.24}$$

assessing the probability of receiving treatment given the individual's characteristics X. Adjusting for PS removes biases associated with included covariates [176]. In addition, by assuming positivity (i.e., all confounder values can be observed in both treated and controls) and no unobserved confounders, the treatment and potential outcomes become independent conditional on $\pi(X_i)$, $T \perp Y(0), Y(1)|\pi(X)$, allowing straightforward effect estimation by matching or regressing the outcome on PS [177].

Another approach, *Inverse Probability Weighting* (IPW), balances the distribution of X across treated and controls by creating a pseudo-population where each original subject is re-weighted using weights:

$$w_i = \frac{T_i}{\pi(X_i)} + \frac{1 - T_i}{1 - \pi(X_i)}. \tag{5.25}$$

The weights can be, for example, incorporated into flexible, even machine-learning-based, outcome models (e.g., weighted regression) to estimate the treatment effect while mitigating selection bias [178].

In our study, focusing on quantifying the effects of OSA ($T = 1$) on $\mathbf{P}$, we employ IPW within the S-learner framework [180]. The S-learner is a baseline approach of meta-learners, enabling flexible estimation of heterogeneous CATE. The S-learner quantifies the outcome using a single model (hence *S*-Learner), including the treatment indicator T as one of its predictors:

$$\mu(x, t) = E[Y^{obs}|X = x, T = t], \tag{5.26}$$

allowing straightforward estimation of CATE from Eq. 5.22 that is easily extrapolated over the entire range of X:

$$\hat{\text{CATE}}(\text{x}) = \hat{\mu}(x, 1) - \hat{\mu}(x, 0). \tag{5.27}$$

For probabilistic outcomes, the Risk-Ratio CATE (RR-CATE) is preferred as it naturally compares the chances of an event:

$$\text{RR-}\hat{\text{C}}\text{ATE}(x) = \frac{\hat{\mu}(x, 1)}{\hat{\mu}(x, 0)}. \tag{5.28}$$

One of the key benefits of S-Learner is its simplicity in extrapolating the (RR-)CATE estimates over and beyond the observed values of X. Unlike other meta-learners (e.g., T- or X-learner [180]) that fit separate response functions for exposed ($T = 1$) and control ($T = 0$) subjects, the S-learner estimates a single model and thus requires less data, while assuming that the effects of the other (non-treatment) variables are shared within groups.

Practical considerations. Care must be taken in interpreting causal effects due to assumptions underlying PS (and so IPW), such as no unobserved confounders and positivity. These

assumptions are challenging to validate rigorously. In summary, addressing confounding is better than ignoring it, but interpretations should consider the assumptions made.

5.2.4 Study use case: effects of OSA on sleep-stage transitions matrix P and derived markers

The practical part of our study links the proposed sleep fingerprint **P** (c.f. Eq. 5.1) and derived markers (c.f., Eq. 5.2-5.13 and Eq. 5.14) to a causal framework for their efficient quantification and estimation of disorder effect. We demonstrate our approach on OSA, the most prevalent sleep disorder and a significant risk factor, and exploit study dataset from BSDB.

To model PS from Eq. 5.24, we applied the logistic regression including confounders the most frequently occurring in the literature: age and gender. Both factors are also known to impact the risk of OSA and at the same time, their value range is not constrained between OSA and healthy subjects, thus meeting the positivity assumption. The PS model included separate predictors of the scaled age above 50 years in decades ($X_{(\text{Age}-50)/10}$), gender indicator ($\mathbb{I}_{\text{male}}$), and their interaction:

$$\pi(X) = P(\text{OSA} = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \mathbb{I}_{\text{male}} + \beta_2 X_{(\text{Age}-50)/10} + \beta_3 \mathbb{I}_{\text{male}} \times X_{(\text{Age}-50)/10})}}. \tag{5.29}$$

The IPW weights based on Eq. 5.25 were used to balance the data concerning the main confounders shared.

To estimate the effects, i.e., (RR)-CATE from Eq. 5.27-5.28, of OSA on the compositional outcome of **P**, the Dirichlet regression, as introduced in Eq. 5.16-5.17, was exploited to model the response within the S-learner framework from Eq. 5.26. Each of the 25 possible transition proportions captured in **P** and indexed as $(i, j) \quad \forall i, j \in \{\text{W, N1, N2, N3, REM}\}$, was modelled using the predictor specific for the corresponding dimension characterized by $\alpha_{(i,j)}$:

$$\begin{aligned}\log(\alpha_{(i,j)}) =& \beta_{(i,j),0} + \beta_{(i,j),1}\mathbb{I}_{\text{male}} + \beta_{(i,j),2}X_{(\text{Age}-50)/10} + \beta_{(i,j),3}\mathbb{I}_{\text{OSA}} + \beta_{(i,j),4}(\mathbb{I}_{\text{OSA}} \times \mathbb{I}_{\text{male}}) + \\ & \beta_{(i,j),5}(\mathbb{I}_{\text{OSA}} \times X_{(\text{AHI}-5)/10}) + \beta_{(i,j),6}(\mathbb{I}_{\text{OSA}} \times \mathbb{I}_{\text{Insomnia\_Com}}) + \beta_{(i,j),7}(\mathbb{I}_{\text{OSA}} \times \mathbb{I}_{\text{NT1\_Com}}) + \\ & \beta_{(i,j),8}(\mathbb{I}_{\text{OSA}} \times \mathbb{I}_{\text{OtherHyp\_Com}}) + \beta_{(i,j),9}(\mathbb{I}_{\text{OSA}} \times \mathbb{I}_{\text{Parasomnia\_Com}}) + \beta_{(i,j),10}(\mathbb{I}_{\text{OSA}} \times \mathbb{I}_{\text{Movement\_Com}}).\end{aligned} \tag{5.30}$$

This log-transformed $\alpha_{(i,j)}$ was regressed on several covariates and interaction terms with a primary goal to separate and quantify the effect of OSA, present as an indicator variable $\mathbb{I}_{\text{OSA}}$. Although this S-learner model was estimated on IPW-balanced data (c.f., Eq. 5.29), the inclusion of age and gender was justified by the necessary adjustment due to their known influence on sleep manifestation. Next, the interaction of OSA with gender was also included, to investigate potential gender-specific phenotypes. In addition, several variables that violating the positivity assumption were included, as they could not be utilized within the PS model due to their disjoint distributions among healthy and OSA subjects. This included the interaction terms of OSA with scaled Apnea Hypopnea Index (AHI), $X_{(\text{AHI}-5)/10}$, denoting number of AHI greater than 5 in tens, capturing the apnea severity as the number of complete or partial breath-arrests per hour. Uniquely, our model adjusts for a comprehensive range of comorbidities present as indicator variables: insomnia ($\mathbb{I}_{\text{Insomnia\_Com}}$), Narcolepsy Type 1 (NT1, $\mathbb{I}_{\text{NT1\_Com}}$), other hypersomnolence except NT1 ($\mathbb{I}_{\text{OtherHyp\_Com}}$), parasomnias ($\mathbb{I}_{\text{Parasomnia\_Com}}$), and movement-related sleep-disorders ($\mathbb{I}_{\text{Movement\_Com}}$). The distribution of AHI and all the comorbidities is completely disjoint, as healthy subjects do not suffer from any disorder/comorbidity and AHI values in OSA subjects are always greater than 5.

To assess uncertainty and calculate confidence intervals (CI) in all strands of our investigations, including the PS model, IPW-balanced S-learner with Dirichlet regression, and subsequent quantification of **P**-derived markers using (RR)-CATE, we implemented a non-parametric bootstrap procedure with 200 repetitions, inspired by [184].

5.3 Results

The main findings of our study are presented in the four subsections:

- *Modelling of sleep-stage transition matrix*, following Figure 5.1a-c, presents the estimation of causal S-learner quantifying the matrix of sleep-stage transition proportions **P**, and the impact of predictors, on IPW-balanced data.
- *Personalized digital markers of sleep dynamics and the effects of OSA*, following Figure 5.1d, introduces principal findings on OSA-markers based on: 1. *raw matrix* **P** exploring the overall prevalence of individual transitions; 2. *derived markers* capturing certain clinical properties by summing up relevant dimensions of **P**; and 3. *derived Markovian matrix* $\mathbf{P}^M$ investigating sleep-stage-specific transition mechanisms related to stage durations. The personalization of markers refers to the estimation of the OSA impact for various levels of age, genders, and apnea-severity, helping to understand how OSA alters sleep and its dynamics across different subpopulations.
- *Predictive Performance of* **P** *Markers on External Data* evaluates the utility of each of the 25 possible sleep-stage transition proportions in identifying subjects with moderate sleep-disordered breathing (AHI > 15). A logistic regression model was trained on the study dataset (BSDB) and applied to the large open-access dataset (SHHS), using only age, gender, and the specific transition proportion as predictors.
- The final part introduces our app, which lets users interactively explore results beyond those shown in this paper (e.g., interactions of OSA with arbitrary comorbidities, evaluation of extreme OSA with AHI$>>$ 30, etc).

5.3.1 Modelling of sleep-stage transition matrix

Propensity score model and IPW balancing

To balance the Berner Sleep Data Base (BSDB) study dataset for the main confounders of gender and age, we used the Inverse Probability Weighting (IPW) strategy, c.f., Figure 5.1a-b. Propensity scores introduced in Eq. 5.24 were used to calculate weights according to Eq. 5.25. The estimates of propensity scores were based on the logistic regression model from Eq. 5.29. The choice of gender and age as the inputs for the IPW was driven by the evidence of existing studies that control for them [101], [106] and clinical evidence that OSA is more prevalent in males and at older ages [149]. In the BSDB exploited, both OSA and healthy subjects can be observed across the entire range of age and genders, thus satisfying the assumption of overlap and positivity [177]. After re-weighting the dataset, the characteristics of age and gender were balanced, which was evidenced by a t-test based on IPW-reweighted means and standard deviations that failed to reject (p-val $>$ 0.05) the null hypothesis of equality of variable means between the OSA and healthy subjects. The weights were subsequently used within the outcome model, enforcing the balanced impact of age and gender, across OSA and healthy subjects.

Outcome model

The proportions of the 25 possible sleep-stage transitions in **P** were modeled using Dirichlet regression (c.f., Figure 5.1c) applied to IPW-balanced data. The model specification followed Eq. 5.30, and the inclusion of the OSA indicator as one of its predictors exploited the causal S-learner framework, enabling a straightforward quantification of (age, gender, apnea-severity)-heterogeneous OSA-effects in terms of Conditional Average Treatment Effect (CATE) and Risk-Ratio CATE (RR-CATE) (c.f., Eq. 5.27-5.28). Simplistically, the CATE and RR-CATE refer to absolute and relative differences between conditioned (i.e., OSA-affected) and control (i.e., healthy) subjects, respectively. The model estimation followed the implementation of Dirichlet regression in R [179]. To assess uncertainty, both in the model coefficients and derived effects, the nonparametric bootstrap with 200 repetitions was used to calculate 95% confidence intervals (CI) based on 2.5% and 97.5% bootstrapped quantiles.

A summary of estimated regression coefficients together with CI for each predictor and transition proportion is provided in Supplementary Table B.1. The estimates indicate a significant influence of both demographics (age and gender), OSA, and its severity (AHI) on sleep-stage dynamics, as at least one of them had a significant impact on each of the transition proportions. The significant interactions of OSA with gender point to the presence of

possible gender-specific OSA phenotypes. The adjustment for comorbidities appears to be essential as the comorbidity indicators influenced most of the transitions.

Given the complex relationship of the marginal effect on the outcome (i.e., transition %'s) with individual coefficients and the actual predictors' value (c.f., Eq. 5.18), we detail results in the intuitive scales of expected percentages, differences (CATE, Eq. 5.27), and risk-ratios (RR-CATE, Eq. 5.28), below.

5.3.2 Personalized digital markers of sleep dynamics and the effects of OSA

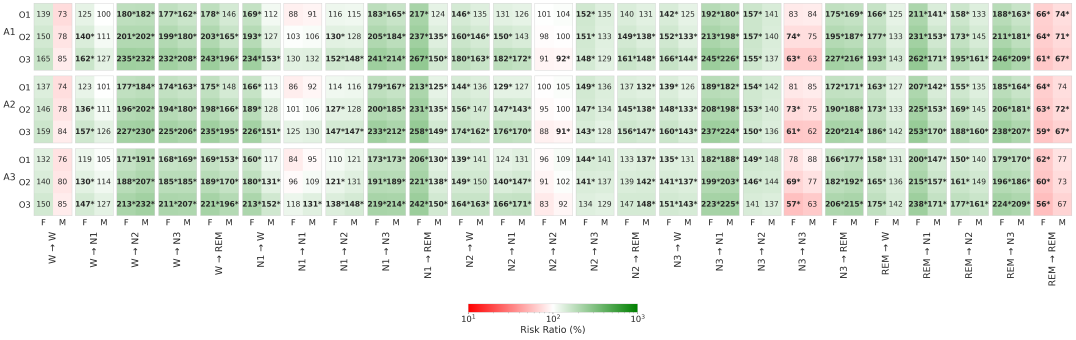
The estimated outcome model enables various scenarios of comparisons of OSA vs healthy, including the raw matrix **P**, derived markers (e.g., % of sleep-stages), and Markovian transition matrix $\mathbf{P}^M$, c.f., Figure 5.1d. All this, for arbitrary values of predictors, provides a wide range of results that can inspire new investigative directions. Since all of our results refer to (possibly derived) transition probabilities (%), we present them in *RR-CATE (CATE)%* format, indicating the amount of *relative (absolute)%* changes, respectively.

Utilizing our model (Eq. 5.30) can extrapolate OSA-effects for arbitrary values of predictors, we showcased the results for three scenarios according to OSA severity, **O1: mild (AHI = 5)**, **O2: moderate (AHI = 15)**, and **O3: severe (AHI = 30)**; three ages: **A1: young (30 years)**, **A2: middle-aged (50 years)**, **A3: older (70 years)**; and for **females (F)** and **males (M)**, ***without comorbidities***. When selecting the most prominent effect in a group, we choose the one according to RR-CATE.

Matrix P of sleep-stage transition proportions

The heatmap in Figure 5.2 shows whether individual transition proportions in **P** (Eq. 5.1) were significantly altered due to specific OSA conditions across different ages and genders. All these aggregated findings are based on detailed results depicted as supplementary heatmap figures showing respective estimates and CI. Specifically, Supplementary Figures B.1 and B.4 depict expected **P** for different ages and OSA-severities for F and M, respectively. Based on that, Supplementary Figures B.2 and B.5 present CATE comparisons between different levels of OSA and healthy individuals of the same demographics, and Supplementary Figures B.3 and B.6 depict the respective RR-CATE.

Figure 5.2: Heatmap of RR-CATE values for OSA effects on sleep-stage transition proportions across gender, age, and OSA severity.



Notes: Risk-Ratio Conditional-Average-Treatment-Effects (RR-CATE) of OSA (compared to a matched healthy subject) on individual dimensions of sleep-fingerprint matrix **P** of sleep-stage transition proportions, per gender (F, M), age (A1, A2, A3), and OSA-severity (O1, O2, O3). Decreased (i.e., RR $<$ 100%) and increased (i.e., RR $>$ 100%) risk-ratios are depicted with red and green shaded backgrounds, respectively. Significant effects are in bold and highlighted with a star (*).

Notably, except for N2 $\rightarrow$ N3 and N3 $\rightarrow$ N2 of A3-F, each significant effect identified for O1 or O2 of both genders was followed with significant effect in the corresponding more severe OSA group. This follows the intuition, that the sleep-stage dynamics and hence also **P** change gradually with increasing prevalence of apnea events (i.e., AHI). The exemption of older F is justified by a significantly lower % of N3, 70.04 (-5.6)% in A3-O3 (c.f., Supplementary Table B.4).

As the entire **P** sums up to 100%, each decrease in a certain proportion is compensated with an increase in one or more other ones. For F, a major decrease is observed in REM $\rightarrow$ REM, with RR-CATE of about 60% across all ages and OSA severities, and the most prominent drop, 55.55 (-4.85)%, in older. This suggests significant REM sleep instability, which could impact cognitive health [185]. The O2- and O3-F also show significantly decreased N3 $\rightarrow$ N3, as low as 57.08 (-6.97)% in A3, indicating disrupted deep-sleep continuity, which may affect physical restoration and memory consolidation [186]. For A1-M, REM $\rightarrow$ REM decreased for all OSA severities, down to 67.5 (-6.19)%, and for A2-(O2,O3), 66.93 (-4.73)%, with the largest declines always in O3. The decreases in all A3-M-OSA groups were not significant, likely due to a larger variance in estimates caused by the limited number of healthy older M in the data. Contrary to F, a decrease in N3 $\rightarrow$ N3 was not significant in M, but a significant decrease in N2 $\rightarrow$ N2 was noted for (A1, A2) O3-OSA, as low as 91.09 (-3.22)%.

For both genders of all ages and OSA severities, several significantly increased transition proportions were identified, distinguishing them from healthy subjects. The most pronounced effects were found in A1-O3-F. The increased W $\rightarrow$ (N2, N3) transitions, up to 234.6 (0.4)%, indicate more frequent arousals attributable to apneic events and subsequent attempts to quickly regain restorative sleep. Increased transitions N1 $\rightarrow$ N3, up to 241.0 (0.4)%, suggest a compensatory mechanism where the body attempts to achieve the restorative effects of deep sleep, bypassing intermediate stages due to frequent sleep disruptions. The increase in N3 $\rightarrow$ (N1, REM) transitions, up to 245.5 (0.3)%, indicates rather infrequent compensatory transitions for reduced N3-continuity, related to a regression to lighter sleep or irregular shifts to REM sleep. Lastly, elevated REM $\rightarrow$ (N1, N3) transitions, up to 261.6 (0.6)%, reflect REM stage instability, with more frequent abrupt changes in sleep depth. Particularly, the atypical transitions between N3 and REM may reflect a build-up of sleep pressure associated with OSA. While such transitions are uncommon under normal conditions, their presence may indicate a compensatory mechanism triggered by long-term disrupted or unrefreshing sleep.

Interestingly, all OSA-F showed a significant increase in awakenings from all sleep stages, (N1, N2, N3, REM) $\rightarrow$ W. For M, there was no increase in REM $\rightarrow$ W in any OSA group, and increases in (N1, N2, N3) $\rightarrow$ W were observed only for O2 and O3. This suggests that in comparison to M, the OSA-F may experience more fragmented sleep due to frequent awakenings from all stages, potentially leading to greater daytime sleepiness, and the presence of insomnia symptoms.

PSG markers derived from P

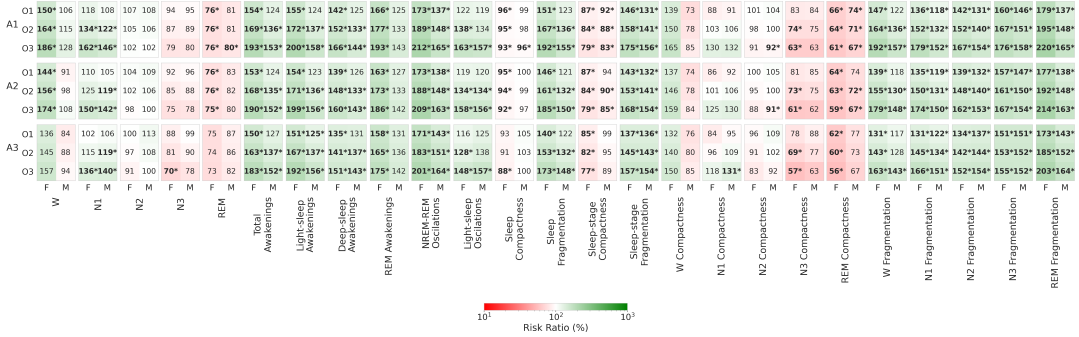
The heatmap in Figure 5.3 aggregates the OSA-effects identified for different PSG markers (c.f., Eq. 5.2-5.13) derived from **P**. Detailed results concerning expected probabilities (%) of their occurrence following Eq. 5.18-5.19, CATE, and RR-CATE for individual age and OSA categories are provided in Supplementary Tables B.2-B.4 for F, and Tables B.5-B.7 for M, respectively.

Regarding the percentagess of individual sleep-stages, the main effect of OSA shared between both genders of all ages is the increase in N1 in O3, with the largest increase of 161.94 (5.53)% in A1-F. The increase affected also all O2-M, up to 122.36 (2.57)% in A1, and A1-O2-F, 134.41 (3.07)%. F seem to have more affected sleep macro-architecture by OSA than M, as for all OSA-severities of (A1, A2)-F an additional increase in W%, up to 185.63% (3.57%) in A1-O3, suggesting a reduced sleep-efficiency, and decreased REM%, as low as 74.9 (-4.25)% in A2-O3, was identified. Except for reduced REM% in A1-O3-M, 79.54 (-4.46)%, these changes were identified only in F.

In addition to increased N3- and REM-awakening from Eq. 5.5-5.6 already discussed above, increased aggregates of total-awakenings (Eq. 5.3), up to 192.55 (2.89)%, and of light-sleep-awakenings (Eq. 5.4), up to 200.35 (2.01)%, were observed in all age and OSA categories with exception of O1-M, with largest effects in A1-O3-F.

A particularly sensitive marker of OSA for all severities appear to be NREM-and-REM oscillations (Eq. 5.7), which were identified as significantly increased across all groups, peaking at 212.48 (3.59)% in A1-O3-F. This marker is elaborated in detail in Figure 5.4 showcasing the expected outcome for F. The upper plots (1a-c) depict the expected probability (%), CATE,

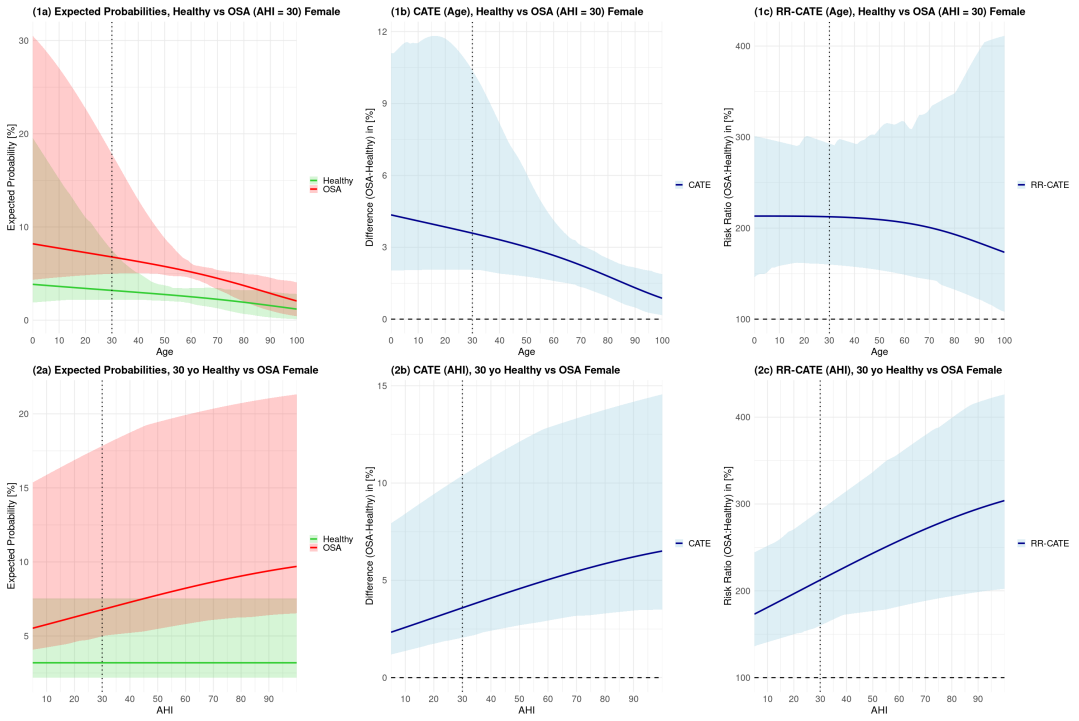
Figure 5.3: Heatmap of RR-CATE values for OSA effects on PSG-markers derived from matrix **P** of sleep-stage transition proportions across gender, age, and OSA severity.



Notes: Risk-Ratio Conditional-Average-Treatment-Effects (RR-CATE) of OSA (compared to a matched healthy subject) on PSG-markers derived from matrix **P** of sleep-stage transition proportions, per gender (F, M), age (A1, A2, A3), and OSA-severity (O1, O2, O3). Decreased (i.e., RR $< 100\%$) and increased (i.e., RR $> 100\%$) risk-ratios are depicted with red and green shaded backgrounds, respectively. Significant effects are in bold and highlighted with a star (*).

and RR-CATE and corresponding CIs for varying age (and fixed AHI), whereas the bottom plots (2a-c) for varying AHI (and fixed age). One can observe, that the effect of OSA remains significant over the entire range of both, age and AHI. The magnitude of the difference tends to decrease with age (c.f., 1b-c), from CATE of about 4.5% in children to 1.5% in older age, likely due to generally shorter sleep with decreasing REM% and lower number of sleep cycles. The effect's size increases rapidly with AHI (c.f., 2b-c), which typically increases with age. The outcomes for M are illustrated in Supplementary Figure B.13.

Figure 5.4: Effects of age and OSA-severities on NREM-REM oscillations, $P(\text{NREM} \rightleftarrows \text{REM})$, in females.



Notes: The left plots (1a, 2a) depict expected probabilities for varying age with fixed AHI = 30, and for varying AHI with fixed age = 30. Based on that, the central (1b, 2b) and right (1c, 2c) plots depict age- and AHI-related CATE and RR-CATE.

Another two highly sensitive derived markers of OSA include sleep- and sleep-stage-fragmentation from Eq. 5.10 and 5.12, referring to probabilities of transitions between wakefulness and sleep, and switching from one non-W stage to the other, respectively. The effect of the sleep-fragmentation was significant across all groups except O1-M and peaked at 192.33 (5.66)% for A1-O3-F. The sleep-stage-fragmentation was increased in all groups, peaking at 174.94 (10.42)% in A1-O3-F. The sleep-stage-fragmentation marker is in-depth elaborated in Supplementary Figures B.14 and B.15, for F and M, respectively.

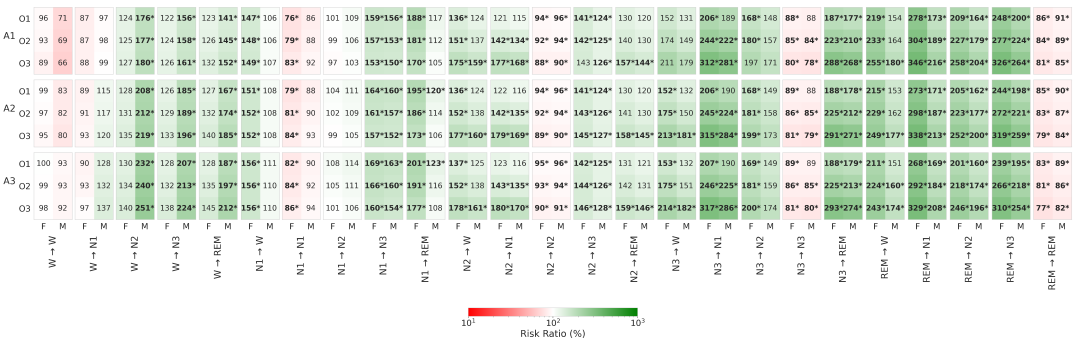
The increased fragmentation is reflected in decreased sleep- and sleep-stage-compactness from Eq. 5.9 and 5.11, referring to staying in not-interrupted sleep and sleep-stage, respectively. Reduced sleep-compactness, down to 88.42 (-9.65)% in A3-O3-F, seems specific to F, suggesting their more frequent apnea-related arousals than M. The sleep-stage-compactness was reduced in all categories of F, down to 76.77 (-16.54)% in A3-O3. This decrease, however, was not present for A3-M and A2-O1-M.

The reduced stage-specific-compactness metrics (e.g., REM $\rightarrow$ REM) were already elaborated in the section on **P**-specific transition %'s. Yet, the stage-specific-fragmentation markers (Eq. 5.13) show significant alterations due to OSA across almost all demographic groups. The only gender-specific difference can be observed in wake-fragmentation, which is increased in all cases of F (likely due to more frequent awakenings experienced), up to 192.1 (2.77)% in A1-O3, but not for O1- and A3-O2-M. The fragmentation related to non-REM (N1, N2, N3) stages increased in all OSA and demographics groups, ranging from 118.29 (1.18)% in N1-fragmentation in A1-O1-M to 178.61% (4.05%) in A1-O3-F. The most pronounced effects were visible in REM-fragmentation, up to 219.51 (2.32)% in A1-O3-F, referring to more than twice as many transitions leaving REM sleep.

Markovian transition matrix $\mathbf{P}^M$ derived from P

Finally, we present the main findings based on $\mathbf{P}^M$, derived from **P** through row normalization as shown in Eq. 5.14. While **P** quantifies the overall probabilities (%) of the 25 sleep-stage transitions, $\mathbf{P}^M$ conditions on the presence of a specific stage, summing to 100% per row. Therefore, whereas **P** evaluates overall chances of observing specific transitions in the hypnogram during the night (e.g., 36.4% of N2 $\rightarrow$ N2 in healthy A1-F), the $\mathbf{P}^M$ evaluates the distribution of the next sleep stage given the current stage (e.g., 84.3% to stay in N2 in healthy A1-F), offering another perspective on the underlying mechanisms of sleep dynamics. The heatmap in Figure 5.5 depicts how individual transitions of $\mathbf{P}^M$ (Eq. 5.14) altered due to specific OSA conditions across different ages and genders. Detailed results on expected transition probabilities of $\mathbf{P}^M$, CATE, and RR-CATE for comparisons of OSA vs healthy are provided in heatmap Supplementary Figures B.7-B.9 and B.10-B.12 for F and M, respectively.

Figure 5.5: Heatmap of RR-CATE values for OSA effects on individual dimensions of row-normalized Markovian transition matrix $\mathbf{P}^M$ across gender, age, and OSA severity.



Notes: Risk-Ratio Conditional-Average-Treatment-Effects (RR-CATE) of OSA (compared to a matched healthy subject) on individual dimensions of row-normalized Markovian transition matrix $\mathbf{P}^M$, per gender (F, M), age (A1, A2, A3), and OSA-severity (O1, O2, O3). Decreased (i.e., RR $<$ 100%) and increased (i.e., RR $>$ 100%) risk-ratios are depicted with red and green shaded backgrounds, respectively. Significant effects are in bold and highlighted with a star (*).

W-transitions: Despite increased occurrences of **P**-transitions from W in F, the respective $\mathbf{P}^M$-dynamic was not significantly altered, indicating that the mechanism of the W-transitions remains similar to healthy subjects, but those transitions tend to occur more often. This suggests that for OSA-F, the overall increased W% is the main trigger of the W-related transitions in **P**. Conversely, M exhibit increased W $\rightarrow$ (N2, N3, REM) transitions, up to 250.5 (1.3)% in A3-O3 for W $\rightarrow$ N2, across all ages and OSA severities, suggesting an increased sleep pressure due to its disruption induced by apneic events.

N1-transitions: Both genders showed increased N1 $\rightarrow$ N3, up to 169.4 (0.9)% in A3-O1-F. Only F experience increased N1 $\rightarrow$ W, up to 156.5 (4.7)% in A3-O1, and decreased N1 $\rightarrow$ N1, as low as 76.5 (-10.3)% in A1-O1. Increased N1 $\rightarrow$ REM transitions were present in all F, up to 201.1 (3.4)% in A3-O1, but only in some of the O1-M, up to 122.7 (1.1)% in A3.

N2-transitions: All groups have decreased N2 $\rightarrow$ N2, down to 88.4 (-9.8)% in A1-O3-F, and, except for A1-O3-F, significantly increased N2 $\rightarrow$ N3 transitions, up to 145.6 (2.2)% in A3-O3-F. All F groups have increased N2 $\rightarrow$ W transitions, up to 177.6 (1.6)% in A3-O3-F, which is present also in all O3-M. N2 $\rightarrow$ N1 increased for all O2 and O3 groups, up to 179.8 (4.2)% in A3-O3-F, and N2 $\rightarrow$ R increased for all O3.

N3-transitions: Across all groups, the N3 dynamic had significantly increased transitions into REM, peaking up to 293.1 (2.1)% in A3-O3-F, pointing to almost three times higher occurrence of these atypical transitions in OSA. Additionally, decreased N3 $\rightarrow$ N3, as low as 77.9 (-18.0)% in A1-O3-M, and increased N3 $\rightarrow$ N1, up to 316.8% (2.5%) in A3-O3-F, were noted for all except O1-M. Transitions N3 $\rightarrow$ W increased in all (A2, A3)-F, up to 214.1% (2.6%) in A3, and only in O3-M of the same demographics.

REM-transitions: The most prominent effects of OSA are visible in changed REM dynamics. The decrease in REM $\rightarrow$ REM in both genders of all ages, down to 77.1 (-20.4)% in A3-O3-F, is compensated by increased transitions into all NREM-stages, up to 345.8 (5.9)% in REM $\rightarrow$ N1 for A1-O3-F. The increased REM $\rightarrow$ W is specific for all F, up to 254.8 (5.3)% in A1-O3-F. For M, these transitions are decreased partially for all O3 and A3-O2, up to 180.0% (2.8%) in A3-O3.

Stage-survival: Finally, following Eq. 5.15, the diagonal elements of $\mathbf{P}^M$ (i.e., probabilities of W $\rightarrow$ W, N1 $\rightarrow$ N1, etc.) simplistically approximate the average expected duration of individual sleep stages, bridging transition dynamics with investigations modelling the sleep-bout durations. Here, naturally, significantly decreased probabilities of staying in a given stage introduced above are equivalent to significantly decreased stage durations.

5.3.3 Predictive Performance of P Markers on External Data

The results of the previous sections focused on quantifying the effects of OSA, specifically explaining how OSA impacts sleep dynamics and its markers. To illustrate the informativeness of these markers, we developed a simple logistic regression model for each of the 25 transition proportions in **P**. The binary outcome variable was defined as moderate sleep-disordered breathing, indicated by AHI > 15, and the predictions were based on three predictors: age, gender, and the percentage of a specific transition. The inclusion of age and gender was motivated by the observed heterogeneity of OSA effects with respect to these factors. Each of these logistic regression models, trained on the study dataset (BSDB), was used to make predictions on the SHHS1 and SHHS2 subsets of SHHS, which contain observational data on subjects who underwent baseline PSG (SHHS1) and follow-up PSG several years later (SHHS2, N = 2,621).

The results in Table 5.2 indicate that each transition proportion included in the simple predictive model demonstrated significant predictive power, as all AUROCs and their confidence intervals were much greater than 50%. The average AUROC performance across proportions was 66.77% for SHHS1 and 65.98% for SHHS2, with standard deviations of 1.79% and 1.82%, respectively, highlighting practical equivalence in generalizations between SHHS1 and SHHS2. This robustness is particularly notable given that (i) we used a simple logistic regression model that assumes monotonic effects of individual predictors and no interactions between them, (ii) we predicted moderate sleep-disordered-breathing (AHI > 15) using only age, gender, and the percentage of a single transition, (iii) the models were trained on a relatively small dataset of 622 patients, and (iv) the generalization was performed from the clinical population of BSDB to the broader public represented by SHHS. All AUROCs are

Table 5.2: AUROC with 95% CI for predicting moderate sleep-disordered breathing from individual sleep-stage transition proportions.

Transition (Predictor)	SHHS1 (N = 5,734)	SHHS2 (N = 2,621)
W → W	67.84* (66.47, 69.21)	66.51* (64.46, 68.55)
W → N1	68.71* (67.36, 70.07)	67.97* (65.95, 69.99)
W → N2	67.04* (65.66, 68.42)	66.03* (63.98, 68.09)
W → N3	66.18* (64.78, 67.57)	65.62* (63.55, 67.69)
W → R	68.42* (67.06, 69.78)	68.47* (66.46, 70.48)
N1 → W	69.28* (67.93, 70.62)	68.39* (66.38, 70.39)
N1 → N1	66.8* (65.42, 68.18)	65.57* (63.51, 67.63)
N1 → N2	66.74* (65.35, 68.12)	66.27* (64.22, 68.33)
N1 → N3	67.2* (65.83, 68.58)	66.23* (64.18, 68.28)
N1 → R	67.54* (66.17, 68.91)	66.86* (64.82, 68.9)
N2 → W	65.18* (63.78, 66.58)	63.79* (61.7, 65.89)
N2 → N1	65.73* (64.34, 67.13)	65.04* (62.97, 67.11)
N2 → N2	59.82* (58.37, 61.28)	59.89* (57.74, 62.05)
N2 → N3	67.43* (66.06, 68.8)	66.46* (64.41, 68.51)
N2 → R	66.06* (64.67, 67.45)	65.16* (63.09, 67.23)
N3 → W	65.82* (64.42, 67.23)	65.12* (63.04, 67.2)
N3 → N1	67.19* (65.82, 68.57)	66.27* (64.22, 68.32)
N3 → N2	67.45* (66.07, 68.82)	66.49* (64.44, 68.54)
N3 → N3	67.89* (66.52, 69.26)	67.61* (65.58, 69.64)
N3 → R	66.94* (65.56, 68.32)	65.57* (63.51, 67.64)
R → W	64.87* (63.47, 66.28)	62.9* (60.79, 65.01)
R → N1	67.45* (66.08, 68.82)	67.15* (65.11, 69.18)
R → N2	66.41* (65.02, 67.8)	66.01* (63.95, 68.06)
R → N3	67.03* (65.65, 68.41)	66.19* (64.14, 68.24)
R → R	68.18* (66.82, 69.54)	67.88* (65.86, 69.89)
Mean ± SD	66.77 ± 1.79	65.98 ± 1.82

Notes: Results are shown for SHHS1 (Sleep Heart Health Study baseline) and SHHS2 (follow-up), of N = number of subjects after exclusion criteria. Mean ± SD (standard deviation) summarizes performance across all transitions. Asterisk (*) denotes significant predictive power with AUROC (> 50%).

very similar, which can be explained by the fact that all transitions in $\mathbf{P}$ are interdependent and numerically share related information.

5.3.4 Interactive R Shiny app

The above-presented results focused on three categories age (30, 50, 70 years), OSA severity (mild, moderate, severe), and both genders, considering a case without sleep-comorbidities. For a deeper exploration of our findings, the volume of which is beyond the scope of this paper, we created a freely accessible app (https://mystatsapps.shinyapps.io/Causal_Sleep_Dynamics/) that interactively displays results for arbitrary values of predictors. As an input, the user specifies the transition(s) of interest by clicking out some of the 25 (5×5) dimensions, age, OSA severity (AHI), and the presence of comorbidities (as indicated in Eq 5.30). Additionally, the user chooses whether CATE and RR-CATE should be displayed for age or AHI (= `CATE-variable`).

As an output, the app displays a total of six panels. The most important one, `Effects of OSA`, displays expected probabilities (%) of selected transitions for healthy vs OSA together with corresponding CATE and RR-CATE. All these outputs are supplemented by 95% CI and are depicted for selected age (range 0-100 years) or AHI (range 5-100), and both genders.

The `Percentual Transition Matrix` and `Markovian Transition Matrix` tabs show the expected matrix of sleep-stage transitions $\mathbf{P}$ and the derived row-normalized $\mathbf{P}^M$ for healthy and OSA subjects of both genders and specified characteristics. In addition, each tab shows matrices of CATE and RR-CATE depicted as heatmaps supplemented with 95% CI.

The `Dirichlet Regression Coefficients` tab summarizes regression coefficients as presented in Supplementary Table B.1. The dimensions of specified transitions of interest from the input are highlighted.

The `Marginal Effects of All Predictors` tab approximate the Eq. 5.18 by calculating the difference in the outcome by a row-indicated change in the predictors' value. The marginal effects that are supplemented with 95% CI are shown concerning four baselines (healthy, OSA) $\times$ (female, male), of specified characteristics from the input. Due to the complex relationship of marginal effect with all Dirichlet dimensions its value changes with the values of predictors (c.f., Eq. 5.18). Hence, their understanding can be particularly useful in understanding the interplay between different levels of demographics, OSA severity, and particularly their interactions with comorbidities, that have been so far understudied.

Finally, the `Sleep Stage Survival` tab depicts survival curves of individual sleep stages, based on diagonal elements on $\mathbf{P}^M$ and Eq. 5.15. Notably, as this quantity is based on the whole-night $\mathbf{P}^M$, survival curves illustrate the overall average duration of individual stages.

5.4 Discussion

Sleep is a complex phenomenon whose finest mechanisms are yet to be fully deciphered. Scoring sleep into a hypnogram of five sleep-wake stages translates it into a simplified, human-readable code, enabling the calculation of PSG markers and their interpretation by clinical personnel. Currently, likely due to non-standardized methodologies and reliance on aggregate counts or summary indices, the representation of sleep-stage dynamics in clinical PSG reports remains limited [8], [187]. Although such markers are reported, they often lack normative values and standardized interpretation guidelines, which may limit their full clinical potential. Yet, existing studies provide strong evidence that more granular characteristics of sleep-stage transitions [98]–[107] or sleep-stage duration/survival [108]–[114] can be specific for various sleep conditions and age. For clinical, economic, and ethical reasons, most of the related research has in common that PSG data were collected in a non-randomised way and were analysed retrospectively, hence subjected to considerable confounding [174]. A minority of studies investigating sleep dynamics addressed confounding either by analyzing subjects with restricted demographic ranges (e.g., [98], [99], [108]), or by selecting typically age- and gender-matched controls (e.g., [100], [102], [104], [114]). This may limit the findings' generalizability or underfit the age- and gender-specific phenotypes.

By exploiting techniques of causal inference (IPW-balancing from Eq. 5.25; S-Learner from Eq. 5.30), our study presents a novel and highly flexible approach to jointly quantify (i) sleep-stage dynamics, (ii) effect of disorder, and (iii) derive several established as well as novel digital markers of sleep. We demonstrate our approach to OSA, the most prevalent sleep condition and a significant risk factor, evidenced to impact sleep macro-strucure and dynamics [103], [106], [109]–[112].

Working with the observational BSDB database, we initially balanced the dataset using IPW-reweighting and addressed the confounding of age and gender, whose distributions differed between healthy and OSA-affected subjects. Ignoring this, it would be challenging to separate the effects of demographics (e.g., of ageing) from OSA, since its prevalence and severity increase with age [112]. To quantify sleep-stage dynamics, we proposed to exploit the matrix $\mathbf{P}$ (Eq. 5.1), consisting of 25 (5×5) interdependent transition proportions. Thanks to the flexibility of $\mathbf{P}$ to quantify all, the dynamics, derived markers, and Markovian $\mathbf{P}^M$, we suggest considering it as a simple *digital sleep-fingerprint*. All dimensions of $\mathbf{P}$ were modelled jointly as an outcome of Dirichlet regression (Eq. 5.17, 5.30), respecting their compositional nature (summing to 100%) and allowing their straightforward aggregation to derive many established and novel PSG markers (c.f., Eq. 5.3-5.13). In contrast, analyzing dependent outcomes, e.g., % of sleep stages and their transitions, separately, such as using (M)ANOVA [106], would lead to biases and disregard constraints on value ranges and cumulative sums. Considering predictors of age and gender allowed outcome model's (Eq. 5.30) adaptation to nonlinear changes in sleep due to ageing and quantification of possible gender phenotypes [18], [139], [166]. Most importantly, the inclusion of the OSA indicator followed the causal S-learner framework [180], allowing direct quantification of OSA effects in terms of CATE and RR-CATE (c.f., Eq. 5.27-5.28) by comparing expected outcomes for healthy individuals of given demographics with hypothetically matched OSA-subject of specified OSA-severity (AHI). Our modelling approach avoids discretization of age and

AHI, and hence allows quantification of personalized (up to OSA-severity and demographics) effects/markers, closely aligning the needs of precision medicine. Even so, it is important to recall that the BSDB dataset contained only 2 cases of paediatric OSA (age < 18 years) and therefore, the conclusions should be taken with care when generalizing them to the pediatric OSA-population. Uniquely, the richness of BSDB allowed us to account for interactions between OSA and several other sleep comorbidities - a clinically well-known and relevant fact (c.f., [10], [188]–[191]), so far either overlooked (e.g., [103], [109]), being admitted but not handled (c.f., [112]), or leading to analysis of subjects with no sleep-comorbidities (e.g., [106], [110]). With 48.6% of OSA subjects in our observational dataset having at least one additional sleep comorbidity, addressing these interactions is crucial for reducing bias and accurately estimating the impact of OSA from other conditions.

The estimated outcome model provides three main dimensions of our results. First, the quantification of sleep fingerprint $\mathbf{P}$ provides information on the % of time spent in individual transitions and compactness of sleep-stages. Several transitions were significantly increased by OSA for all demographics and AHI-severity groups: W $\rightarrow$ (N2, N3), N1 $\rightarrow$ N3, N3 $\rightarrow$ (N1, REM), and REM $\rightarrow$ (N1, N3), all peaking with RR-CATE >200%. Despite their rare presence in healthy subjects, our findings suggest they may be a sensitive marker of OSA. In addition, all OSA-F had significantly increased (N1, N2, N3, REM) $\rightarrow$ W, W $\rightarrow$ REM, N1 $\rightarrow$ REM, REM $\rightarrow$ (W, N2), and decreased REM $\rightarrow$ REM, suggesting their higher vulnerability to awakenings and REM-disruptions in comparison to M, for whom these effects were observed only partially. This finding may also be linked to more likely REM-OSA in F [192]. These results suggest that female OSA patients may experience subtler forms of sleep disruption, such as increased REM instability and awakenings, which could contribute to the under-recognition of OSA burden in women if relying solely on oxygen desaturation metrics. Secondly, by aggregating dimensions of $\mathbf{P}$, one can derive standard PSG markers (e.q., % of sleep-stages), and many novel proposed ones, that may be specific to particular conditions. For all demographic and AHI groups, OSA significantly increased NREM-REM oscillations (c.f., Eq. 5.7), overall sleep-stage fragmentation (c.f., Eq. 5.12), and (N1, N2, N3, REM)-specific fragmentations (c.f., Eq. 5.13). In addition, all, sleep-, light-sleep, and deep-sleep-awakenings (c.f., Eq. 5.3-5.5), were increased for all moderate and severe-OSA groups. Finally, row-normalizing $\mathbf{P}$ yields the Markovian $\mathbf{P}^M$, which quantifies the probabilistic distribution of the next phase given the current state, thus investigating deeper dynamic mechanisms. For all age and AHI groups, OSA increased N1 $\rightarrow$ N3, N3 $\rightarrow$ REM, REM $\rightarrow$ (N1, N2, N3), and decreased REM $\rightarrow$ REM and N2 $\rightarrow$ N2. All moderate and severe OSA had also increased N3 $\rightarrow$ N1 and decreased N3 $\rightarrow$ N3. For all OSA-M, an additional increase in W $\rightarrow$ (N2, N3, REM) and for all OSA-F increase in N1 $\rightarrow$ (W, REM), (N2, REM) $\rightarrow$ W and decreased N1 $\rightarrow$ N1 was observed. Furthermore, we demonstrated that $\mathbf{P}^M$ can also be used to model sleep-stage survival (Eq. 5.15), bridging the two principal directions of sleep dynamics research: sleep-transitions [98]–[107] and sleep-stage bout duration quantification [108]–[114]. The merit of the stage survival analysis includes the evaluation of the functional form of the distribution. We can learn their statistical property which provides insights into the underlying mechanism.

To underscore the diagnostic utility of our findings, we evaluated the predictive power of individual transition proportions in $\mathbf{P}$ on external data from SHHS, containing a broad population of subjects from the general public. For each transition, we developed a simple logistic regression model using age, gender, and the specific transition percentage as predictors, and assessed its ability to identify moderate sleep-disordered breathing (AHI > 15). Results showed significant predictive utility across all 25 transitions, with all AUROC values exceeding 50% (range of 59.82-69.28), and their average of 66.77% for SHHS1 and 65.98% for SHHS2, with respective standard deviations of 1.79% and 1.82%. This robust performance highlights the generalizability of the derived markers from the BSDB dataset to a broader population while confirming the informativeness of individual transitions as predictors of sleep-disordered breathing. Higher predictive performance can be expected when including additional predictors not reflected in $\mathbf{P}$ (e.g., total-sleep-time, sleep-latency), their interactions with specific proportion, using all proportions jointly, or using a more complex predictive model than logistic regression.

5.5 Conclusion

In summary, our findings from different perspectives confirm that OSA is associated with reduced continuity of N2, N3, and REM sleep, reflected by increased sleep fragmentation [103], [106], [109]–[112] at both the conventional sleep-to-wake level and in the proposed markers of stage-specific dynamics. By exploiting the matrices $\mathbf{P}$ and $\mathbf{P}^M$, we identified OSA-specific transitions contributing to these alterations, particularly atypical transitions from light to deep sleep and oscillations between N3 and REM. These transitions, though rare in healthy individuals, may serve as sensitive markers of OSA, possibly reflecting compensatory mechanisms where the body attempts to regain restorative states, either after their frequent disruption by apneic events or following long-term accumulation of disrupted and unrefreshing sleep. These findings contribute to growing evidence that OSA phenotypes vary across demographic groups and may benefit from personalized clinical interpretation. Our results suggest that females with OSA exhibit increased REM-stage instability and a higher frequency of awakenings from multiple sleep stages, despite often presenting with milder oxygen desaturation—patterns that may elude detection by AHI alone when compared to males [193]–[195]. This aligns with prior reports of REM- or arousal-dominant OSA profiles in women, often accompanied by insomnia-like symptoms [18], [139]. By quantifying stage-specific dynamics, our framework may support more refined diagnostic stratification and treatment decisions—especially in subgroups historically underrepresented or mischaracterized by standard PSG indices. For instance, women with OSA—who often present with insomnia-like symptoms and have a higher risk of comorbid depression—may benefit from personalized treatment approaches that contextualize available markers, such as REM-related instability, frequent awakenings, or shorter apneas [195], to more precisely identify the plausible contributing factors—be it OSA, insomnia, depression, or others.. This may guide the use of CPAP with cognitive behavioral therapy for insomnia (CBT-I), or considering oral appliance therapy (OAT) in milder-AHI cases [194], [195]. The results of our work are also available as an interactive app, allowing in-depth exploration of results and proposed markers for arbitrary demographics, OSA severity, and their interactions with other sleep comorbidities.

Our approach to support diagnostics, has broader applicability beyond the OSA use-case, as sleep dynamics and their markers can be specific to other sleep disorders, such as narcolepsy, insomnia, periodic limb movement disorder, and others. With the rise of telemedicine and increasing use of wearables, investigating sleep dynamics and its markers could become a valuable screening tool for assessing the risk of psychiatric (e.g., depression, schizophrenia, etc.) and neurodegenerative disorders (e.g., Parkinson's disorder, Alzheimer's disease, etc.), which are evidenced to be associated with disrupted sleep [3], [6], [33]. Even though the consumer devices provide - compared to clinical PSG - lower quality signals and hypnograms, adaptation and re-estimation of our approach to these data has still great potential to provide valuable insights. Furthermore, with advances in automatic sleep-scoring tools that offer hypnodensity beyond the standard hypnogram [61], our framework could enhance the understanding of sleep micro-events and more granular sleep dynamics, when hypnograms on less than 30-second windows would be used as data for our model.

5.6 Limitations

Our future work will extend our approach to address several of its limitations. Following the ideas of Schlemmer et al. (2015) [105], we aim to extend it to the second-order sleep-stage transitions that would require quantifying a 125 (= 5 x 25) dimensional transition cube. Next, we plan to account for time spent asleep and investigate dynamics at different times of the night. Currently, we have focused on transitions aggregated over the entire sleep period, but recognizing the non-stationary nature of sleep offers opportunities for identifying even more specific markers. This would also concern the quantification of sleep-stage survival or duration, which our current work approximated by an overall night expectation. Additionally, we plan to consider whether the subject's apnea events are REM- or NREM-dominant, which may reveal additional phenotypes, and to assess whether the proposed markers can also help distinguish obstructive sleep apnea (OSA) from central sleep apnea (CSA). Further,

we plan to investigate in greater detail the interaction of OSA with comorbidities, which can already be explored in our app. Finally, our current framework captures sleep dynamics at the macro-structural (sleep-stage) level, relying on stage annotations from standard hypnograms [8]. It does not directly account for EEG-based micro-events such as brief arousals or microstructural instability captured by the Cyclic Alternating Pattern [196]–[200], which has shown clinical relevance in characterizing sleep instability, evaluating treatment response (e.g., to CPAP), and supporting diagnostics. With the rise of automatic sleep-scoring algorithms that produce hypnodensity outputs [61]—i.e., probabilistic stage predictions at sub-30-second resolution—there is growing potential to adapt our framework to this finer temporal scale. The domains of sleep-stage and micro-structural dynamics can now be seen as complementary, and our future work will aim to bridge them.

Chapter 6

Unveiling Sleep Dysregulation in Chronic Fatigue Syndrome with and without Fibromyalgia Through Bayesian Networks

Abstract

Chronic Fatigue Syndrome (CFS) and Fibromyalgia (FM) often co-occur as medically unexplained conditions linked to disrupted physiological regulation, including altered sleep. Building on the work of Kishi et al. [201], who identified differences in sleep-stage transitions in women with CFS and CFS+FM, we exploited the same strictly controlled clinical cohort using a Bayesian Network (BN) to quantify detailed patterns of sleep and its dynamics. Our BN confirmed that sleep transitions are best described as a second-order process [107], achieving a next-stage predictive accuracy of 70.6%, validated on two independent data sets with domain shifts (60.1–69.8% accuracy). Notably, we demonstrated that sleep dynamics can reveal the actual diagnoses. Our BN successfully differentiated healthy, CFS, and CFS+FM individuals, achieving an AUROC of 75.4%. Using interventions, we quantified sleep alterations attributable specifically to CFS and CFS+FM, identifying changes in stage prevalence, durations, and first- and second-order transitions. These findings reveal novel markers for CFS and CFS+FM in early-to-mid-adulthood women, offering insights into their physiological mechanisms and supporting their clinical differentiation.

Keywords:
Chronic Fatigue Syndrome, Fibromyalgia, Sleep Dynamics, Polysomnography, Bayesian Network

6.1 Introduction

Chronic Fatigue Syndrome (CFS) and Fibromyalgia (FM) co-occur in up to 70% of cases [202]. These conditions share symptoms such as disrupted sleep and exhaustion but have distinct clinical profiles: CFS is characterized by severe, unexplained fatigue worsened by exertion [203], whereas FM is defined by widespread musculoskeletal pain and sensory hypersensitivity [204]. Both conditions disproportionately affect females, with prevalence up to four times higher than in males [205], and are most commonly reported in young to middle-aged adults [203], [204], [206]. They are frequently accompanied by other clinical conditions, including psychiatric and specific sleep disorders [207], [208], complicating the quantification of their underlying effects. Consequently, clinical reviews of existing - mostly observational - studies often lack evidence of their systematic impacts on sleep architecture [208].

The study cohort by Kishi et al. [201] minimized confounding factors and collected polysomnographic (PSG) data from a strictly controlled set of healthy (H), CFS, and CFS+FM women aged 25–55. Exploratory data analysis revealed changes in sleep stage durations and proportions and identified first-order transitions as potential markers enabling clinical interpretation of physiological dysregulation in CFS and CFS+FM.

Recent research in individuals with or without sleep disorders showed that sleep-stage transitions are optimally modelled and analyzed as a second-order process [105], [107]. Leveraging these insights and the CFS/FM dataset [201], we (i) *implement a Bayesian Network (BN) capable of both next-stage prediction and diagnostics*, (ii) *validate the second-order optimality even in a clinical cohort*, and based on that (iii) *identify novel markers for CFS and CFS+FM based on two-stage transitions*, providing novel insights into their physiology and supporting their clinical differentiation.

6.2 Materials and Methods

6.2.1 Data

Primary Cohort. The data from [201] comprises PSG recordings from 52 women, carefully selected to ensure homogeneity and avoid confounding. The cohort included 26 healthy controls (H, aged 38 $\pm$ 8 years), 14 individuals with CFS only (aged 37 $\pm$ 9 years), and 12 individuals with CFS and FM (CFS+FM) (age: 41 $\pm$ 6 years). Rigorous exclusion criteria were applied, including the presence of clinically evident sleep disorders or other psychiatric conditions. Subjects also refrained from alcohol, caffeine and strenuous activities before the study, and menstruating individuals were evaluated during the follicular phase of their cycles. The PSG data were recorded during a single night in a controlled hospital environment, with sleep stages scored every 30 seconds. This carefully curated data set enables robust estimation of the underlying effects of CFS and CFS+FM in early to mid-adulthood women.

Validation Cohorts. The *Bern Sleep–Wake Registry* (BSWR) from the University Hospital Bern and the open-access *Sleep Heart Health Study* (SHHS) are clinical and general-population data sets used to assess the robustness and validate the next-stage predictions of our developed model. To ensure demographic alignment with the primary cohort, subsets of 834 and 1227 women aged 20–60 were selected from the BSWR and baseline-SHHS (SHHS1), respectively. The BSWR challenged the model's predictive capabilities with a population of sleep-disordered subjects, while SHHS1 assessed it in a general population.

To ensure consistency across analyses, sleep-scoring in all data sets was standardized to five sleep-wake stages following the AASM guidelines [8]: W = Wake, R = Rapid-eye-movement sleep, and (N1, N2, N3) non-R sleep-states.

Preprocessing. Having a controlled homogeneous study population (women of the same age), we considered Health Status (HS): H, CFS, CFS+FM, as the only demographic variable. When modelling sleep dynamics, we ignored the PSG recordings before the first non-W stage. We identified continuous bouts (runs) of each stage—denoted S_t, indexed by t—and recorded their durations (D_t). This reduced the original 44,581 sleep-stage-epochs to 7,254 bouts. For each bout, we also recorded the time-since-sleep-onset (T_t, TSSO) and cumulative characteristics (C_t) monitoring either sleep-time (CST=N1+N2+N3+R) or restorative-sleep-time (CRST=N3+R). To utilize the existing Bayesian inference implementation [209], we discretized the TSSO into five 90 minutes categories ($<$90, 90-180,..., $>$360) of expected sleep cycles and split C_t and D_t variables into four groups based on (25, 50, 75)%-quantiles, with the possible additional class of 0, if present in the corresponding variable. The validation cohorts underwent the same preprocessing, yielding 113,071 and 150,296 bouts, respectively.

6.2.2 Bayesian Networks to Capture Sleep Stage Dynamics

A Bayesian Network (BN) is a statistical framework that encodes probabilistic relationships between variables and can represent cause-effect relationships under additional causal assumptions [209]. These relationships can be learned from data (structure learning), defined by experts (incorporating domain knowledge), or by combining both approaches. Represented as a Directed Acyclic Graph (DAG), BNs offer several advantages, including reduced parameter complexity and interpretable predictions—a critical requirement in the healthcare field.

A compelling feature of BNs is their ability to fix specific nodes (variables) at desired levels, such as the health status (HS) to H or CFS, representing what is referred to as an *intervention*. This enables do-calculus and the simulation of causal counterfactuals [210], [211], addressing what-if questions such as ours of *how sleep patterns change if a healthy individual were to develop CFS or CFS+FM*. Dynamic BNs (DBNs) extend the approach to temporal processes by incorporating dependencies across time, including lagged features. This makes them particularly suited for modelling sleep transitions.

Experimental Setup. Rather than relying only on data-driven structure learning algorithms, we predefined dependencies using expert knowledge, as shown in Figure 1. This included mandatory (solid) edges encoding the impact of all previous stages on the following ones (in red) and the impact of HS on S, C, D (in green). The possible (dashed) impact of TSSO (T) on S, C, D (in yellow) was also considered. Further, we hypothesized that transitions in S might be better explained by considering cumulative sleep variables C (in orange), naturally depending on T. Existing work suggested that including stage-duration D, depending on S (in red), might boost next-stage predictions (in blue) [107]. To systematically evaluate each variable's inclusion and identify the optimal structure, including BN lag/order (0–4), we fitted the BN for each possible combination of non-mandatory nodes and their associated dependencies (edges), and used linear regression to associate BN performance (for next-stage and HS prediction) with indicators of each variable's inclusion. Clinically, beyond quantifying the effects of CFS and CFS+FM (HS-node), this allowed us to test whether the cumulative sleep (CST, CRST) better explains sleep dynamics than TSSO. Despite the well-known effect of TSSO on sleep macro-architecture (e.g., higher R% in the second half of the night), its influence on dynamics remains inconclusive [107].

6.3 Results

6.3.1 Descriptive Statistics

Traditional sleep variables of the primary dataset are described in detail in the original work, which also reports their Tukey-Kramer multiple comparisons concerning the HS [201]. The significant differences were identified for the total sleep time [mins] (H > CFS), N1 and N2 [mins] (H > CFS+FM), N3 [mins] (CFS+FM > H), and REM [mins] (H > CFS), c.f., Table 2 in [201].

Occurence of stage-specific bouts and their duration is presented in Table 6.1. H experienced significantly more R-bouts than both CFS conditions and more N1-bouts than CFS+FM. In addition, CFS+FM exhibit more N3-bouts than CFS only. The subject-aggregated means of stage-specific bout durations did not exhibit significant differences across HS.

Table 6.1: Mean (SD) bout statistics for healthy (H), CFS, and CFS+FM subjects.

Stage	Characteristic	H	CFS	CFS+FM	Significant Pairs
W	Bouts	22.5 (6.4)	22.5 (8.5)	19.7 (5.6)	-
	Duration	2.4 (1.5)	3.6 (3)	3.2 (2.1)	-
N1	Bouts	44.6 (17.2)	37.5 (17.7)	29.8 (8.9)	H - (CFS+FM)*
	Duration	1 (0.2)	1 (0.2)	0.9 (0.2)	-
N2	Bouts	50.2 (16)	43.1 (11.9)	52.7 (14.8)	-
	Duration	5 (2)	5.1 (1.7)	4.1 (2.1)	-
N3	Bouts	18.4 (10.2)	15.6 (6.6)	25.1 (12)	CFS - (CFS+FM)*
	Duration	2.2 (2.2)	2.8 (1.9)	3.3 (2)	-
R	Bouts	13.4 (7)	6.5 (4.8)	8.2 (3.8)	H - CFS**; H - (CFS+FM)*
	Duration	8.2 (5.8)	12.6 (8.5)	10.8 (7.9)	-

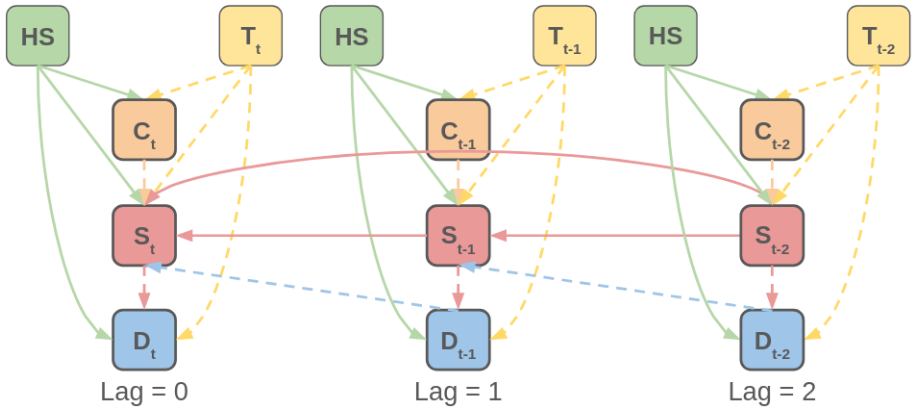
Notes: Bouts indicate the average number of stage runs, and Duration their mean length in minutes. Significant pairwise comparisons according to the Tukey-Kramer procedure are marked with * and ** for p-value < (0.05 and 0.01), respectively.

6.3.2 Structure Identification

The structure of BN was selected based on the computational experiment described above, testing all expertly-predefined node combinations from Figure 6.1 under restricted settings of temporal ordering and HS being the underlying cause of transitions. This evaluation

used HS-balanced 3-fold cross-validation (CV) with subject-wise splits, allowing performance quantification for each variable combination while ensuring a reasonable number of subjects were included in the testing fold.

Figure 6.1: Full-structure Bayesian network with lag = 2.



Notes: HS = health status (healthy, CFS, or CFS + FM); T_t = time since sleep onset; C_t = cumulative sleep; S_t = sleep stage; D_t = sleep-stage duration, chronologically indexed by t.

The performance metrics used included: **next-stage accuracy** and **F1-score**, and **average AUROC** (of AUROCs specific to H, CFS, and CFS+FM). The results are summarized in Table 6.2.

Table 6.2: The impact of BN-included variables on the performance metrics.

Variable	Accuracy [S_t]	F1-score [S_t]	AUROC [HS]
lag = 0	44.00	50.84	71.59
lag = 1	68.03	72.75	74.13
lag = 2	72.08	73.07	74.29
lag = 3	68.91	70.05	75.63
lag = 4	64.78	65.94	75.85
TSSO	-4.14	-4.62	-4.14
Stage-Duration	-3.29	-3.87	2.45
CST	-1.53	-1.65	-3.98
CRST	-12.03	-12.56	-8.42
Model's F(9, 51)	1252.09	2323.33	4715.47
Model's R^2_{adjusted}	0.995	0.997	0.999

Notes: Significant variable associations and model explanations based on F-test are highlighted as p-value < 0.05 , 0.01 , and 0.001 , respectively. The R^2_{adjusted} (not tested) and F-statistic refer to regression models evaluating the systematic impact of included variables on the performance metric across different BN-settings and not to any specific BN.

Based on next-stage performance metrics, we identified lag=2 as optimal, confirming [107], as both accuracy and F1-score were the highest and appear to decrease with larger lags. The AUROC, indicating capability to identify HS, was up to 1.56% better for higher lags, but their consideration would lead to an expected decrease of up to 7.3% in accuracy/F1-score. All TSSO, CST, and CRST yielded a systematic decrease in all performance metrics. This may suggest that sleep dynamics and HS identification are either unrelated to these variables or that the BN was under their inclusions over-parametrized, as the number of parameters to predict the S_t just from S_{t-1}, S_{t-2}, HS involves $75 = 5{\times}5{\times}3$ parameters which scale by 4-5 with inclusion of every additional (TSSO, CST, CRST) variable. Based on that, we chose the BN of lag = 2 with included stage durations as the final model to demonstrate the CFS and CFS+FM effects. Despite slightly reduced next-stage predictive accuracy due to duration inclusion, this model seems to significantly enhance the identification of HS. Our evaluations tried to find a compromise between the best performance in the next stage and diagnosis identifications.

6.3.3 Performance and Generalization

The final BN (lag = 2, including stage durations) achieved 70.61 (1.9)% and 69.2 (2.7)% in mean (SD) on-subject next-stage accuracy and F1-score, and the HS AUROC of 75.36 (8.3)%. For each subject, we estimated HS probabilities by averaging posterior queries over all triplets of sleep stages and durations.

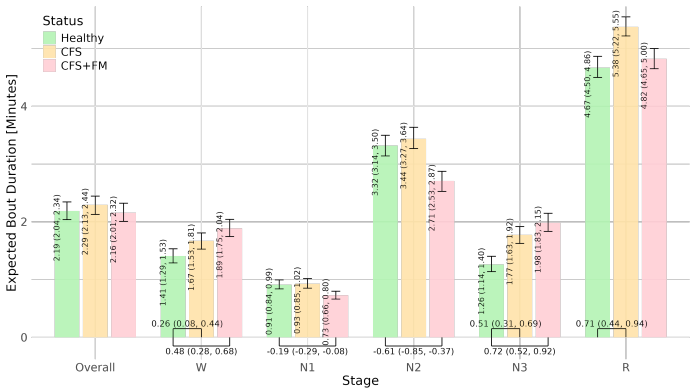
To further test the robustness of the final BN to capture sleep dynamics, we evaluated its predictive accuracy on BSWR and SHHS1. Despite training on a small sample of 52 strictly controlled subjects, BN achieved 69.78 (7.25)% and 60.1 (11.62)% in mean (SD) on-subject accuracy, 70.94 (9.1)% and 59.83 (11.56)% in on-subject F1-score, on BSWR and SHHS1, respectively. Considering that both test data sets represent out-of-domain samples from general and clinical cohorts, respectively, with considerable domain shifts, these results suggest the particular robustness of our BN. In contrast, similar work reported 62.2% testing accuracy (corresponded to in-domain cross-validation assessment) on a broad sample of 3,202 PSG recordings with excluded sleep-disorders [107].

6.3.4 Effects of CFS and CFS+FM via Interventions

We evaluated three interventions by fixing the HS node of our final BN to H, CFS, and CFS+FM levels, allowing sampling from arbitrary nodes under specified conditions. Assuming no hidden confounding, which is reasonable in our strictly controlled cohort, comparing samples for CFS-vs-H and (CFS+FM)-vs-H enables estimating the causal effects of the two conditions. Arbitrary 95% credible intervals (CI) were constructed by generating 1,000×1,000 samples and calculating median (= estimate) and (2.5, 97.5)%-quantiles (= CI-bounds).

Bouts Duration: Figure 6.2 presents BN-based CIs for expected stage durations. Discretized D_t levels were represented by mid-points and multiplied by obtained samples. Both CFS and CFS+FM exhibit prolonged W and N3 durations, indicating reduced sleep efficiency and increased physically-restorative drive. CFS additionally exhibits extended R stages, linked to cognitive restoration, despite fewer R bouts. In contrast, CFS+FM shows shorter N1 durations, likely compensating for increased W and N3. Notably, CFS does not display reduced durations in any stage, suggesting compact sleep despite decreased efficiency.

Figure 6.2: Expected durations of sleep-stage bouts for H, CFS, and CFS+FM groups.

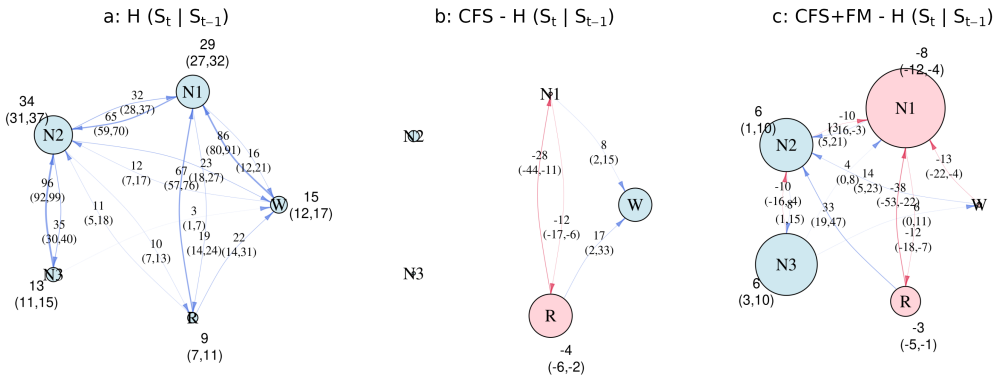


Notes: Durations are shown with 95% confidence intervals (CIs) as vertical error bars. Horizontal brackets indicate significant between-group differences, reported with their estimates and 95% CIs.

First-order transitions $(S_t \mid S_{t-1})$ expected for H are shown in Figure 6.3.a and the CFS and CFS+FM effects in Figure 6.3.(b-c). The effects were quantified without conditioning on any particular stage and describe the overall sleep dynamics. Below, we write in **bold** alterations by at least 10%. CFS showed reduced R% and increased N1→W, **R→W** that were

compensated by decreased **N1⇄R**. The changes were more pronounced in CFS+FM, which showed increased (N2, N3)% and decreased (N1, R)%. Further, CFS+FM exhibited significantly increased (W, **N1**, **R**)→**N2**, N2→N3, N3→(W, N1), and decreased **(W, N2, R)→N1**, **N1→R**, and **N3→N2**. Our findings confirm all alterations found by [201] in their Figure 1. We additionally identified increased N1→W (c.f., [102]) in CFS (compensation for decreased N1→R) and disruptions in N2 for CFS+FM [201].

Figure 6.3: Lag-1 sleep-stage transition dynamics for Healthy (H), Chronic Fatigue Syndrome (CFS), and CFS with Fibromyalgia (CFS+FM).



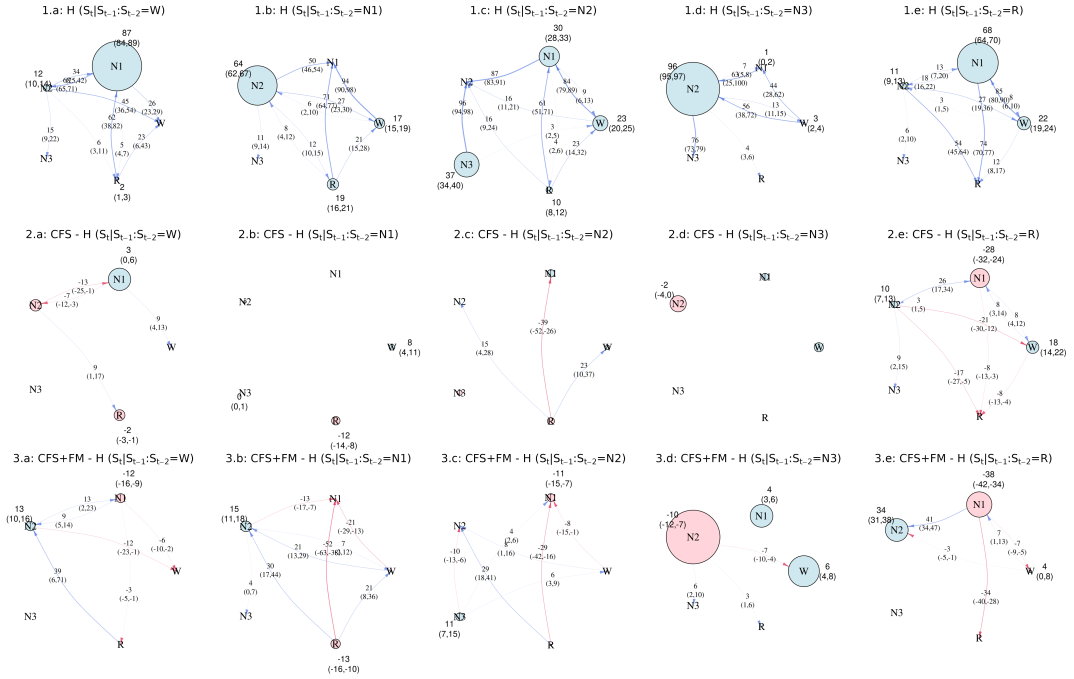
Notes: Panel (a) illustrates the expected transitions for H, with node sizes proportional to the prevalence of S_{t-1} stages and edges indicating transition $(S_t \mid S_{t-1})$ probabilities. Panels (b) and (c) depict the differences in stage prevalence and transition probabilities due to CFS and CFS+FM, in comparison to H, respectively. Positive and negative values are shown in blue and red, respectively, and significant alterations are annotated with their estimates and 95% credible intervals.

Second-order transitions $(S_t \mid S_{t-1}, S_{t-2})$ in Figure 6.4 provide deeper insights into sleep dynamics. The first row represents expected transitions for H, while rows 2 and 3 depict the effects of CFS and CFS+FM. Each column (a–e) corresponds to a different starting stage S_{t-2}. In some cases, the alterations are only in S_{t-1} (nodes), or follow-up transitions (S_{t-1}→S_t, edges), both conditioned on S_{t-2} and extending the unconditioned first-order results from Figure 6.3.

In ***CFS***, key disruptions included increased **R→W** (with subsequent increases in N1 and decreases in R) and **R→N2** (followed by increased **N1**, N3, and decreased **W**, **R**), along with reduced **R→N1**. These patterns suggest an impaired ability to achieve or maintain restorative R sleep, compensated by non-restorative transitions within light sleep (N1, N2). Additionally, W→R, common in healthy individuals during the second half of the night, was decreased and supplemented by W→N1. More frequent N1→W, at the expense of **N1→R**, further contributed to reduced sleep-efficiency and increased fragmentation.

In ***CFS+FM***, disruptions included increased R→W (with subsequent increases in N1 and decreases in N2) and **R→N2**, along with reduced **R→N1** (followed by decreased W and **R**, and increased **N2**). Particularly increased **N2→N3** (followed by increased transitions to W and N1, and reduced to **N2**), reflecting a compensatory drive for deep sleep (N3) likely linked to FM's restorative needs, while also indicating difficulty maintaining smooth sleep cycling. Reduced **W→N1** and increased **W→N2** suggest a shift towards intermediate sleep stages at the expense of lighter sleep, possibly as a response to pain-related disruptions. Increased awakenings from N3 (compensated by reduced **N3→N2**) and from R further destabilized transitions between restorative and lighter stages, amplifying sleep fragmentation and reducing efficiency. These findings align with FM's symptomatology, where widespread pain increases the need for deep sleep (N3) but disrupts restorative sleep transitions, highlighting the need for tailored treatments to improve both sleep and pain management.

Figure 6.4: Lag-2 sleep-stage transition dynamics for Healthy (H), Chronic Fatigue Syndrome (CFS), and CFS with Fibromyalgia (CFS+FM).



Notes: The first row shows the expected dynamics for H, while rows 2 and 3 display their changes due to CFS and CFS+FM, in comparison to H, respectively. Node sizes represent S_{t-1} prevalence (or its difference), and edges illustrate transition probabilities $(S_t \mid S_{t-1})$ or their differences, both conditioned on S_{t-2}. Positive and negative values are shown in blue and red, respectively, and significant alterations are annotated with estimates and 95% credible intervals.

6.4 Discussion

In this study, we constructed a Bayesian Network (BN) to quantify the effects of Chronic Fatigue Syndrome (CFS) and its interaction with Fibromyalgia (FM) on sleep dynamics. Using a strictly controlled dataset [201], we confirmed that second-order transitions $(S_t \mid S_{t-1}, S_{t-2})$ optimally describe sleep patterns, extending findings from non-clinical populations [107]. Despite a relatively small dataset of 7,254 bouts from 52 subjects, our BN achieved robust next-stage predictions with in-domain (out-of-domain) accuracies of 70.6% (60.1–69.8%), respectively. This capability enabled the successful differentiation of healthy (H), CFS, and CFS+FM groups (AUROC: 75.4%), showcasing sleep dynamics' potential for diagnostics. Based on that, we used interventions to quantify the effects of CFS and CFS+FM compared to H on different aspects of sleep dynamics.

Both conditions exhibited prolonged wakefulness (W) and N3 stages, reflecting reduced sleep efficiency (aligning with insomnia-symptoms in CFS [207]) and increased physical restoration needs, particularly pronounced in CFS+FM. Additionally, CFS showed extended R durations related to an increased sympathetic activity and a higher need for cognitive restoration, while CFS+FM demonstrated reduced durations of N1 and N2. Interestingly, the duration of any stage did not decrease in CFS, suggesting that their sleep - despite reduced efficiency - may remain relatively compact.

First-order transitions confirmed all previous findings [201], and - thanks to the joint estimation of transition-probabilities in our BN (as opposed to the pairwise comparisons in [201]), revealed three additional compensatory transitions. CFS is marked by frequent and prolonged awakenings from the N1 and R stages, disrupting "healthy" oscillations between them. This suggests reduced sleep efficiency at the expense of R sleep, potentially contributing to fatigue from both, insufficient sleep quantity and inadequate autonomic or cognitive restoration. In contrast, CFS+FM is characterized by awakenings from deep N3 sleep, into which they tend to transition more frequently. FM, associated with physical pain and discomfort [204], appears to drive both the increased N3 duration and the pressure to transition to N2 instead of N3 across stages.

The second-order transitions provided a novel and detailed perspective on sleep alterations. Both conditions exhibited increased transitions into W, particularly from R, reflecting reduced sleep efficiency. For CFS, fewer alterations were observed, consistent with their longer bouts. The results highlighted CFS-specific patterns of awakenings from N1 and R, difficulties maintaining R (due to transitions into N2), and challenges achieving R. These disruptions may represent the patients' common complaint of "unrefreshing sleep", either as a cause or a consequence of fatigue, as commonly reported in CFS. In contrast, CFS+FM showed more widespread alterations, including frequent awakenings from both R and N3, coupled with a marked compensatory drive to achieve and sustain N3, likely driven by the physical symptoms of FM.

6.5 Conclusion

Our study confirms that sleep transitions are best described as a second-order process, even in diseased clinical subjects. Using a strictly controlled cohort of young-to-middle-aged women, we identified the effects of CFS and CFS+FM on alteration sleep and its dynamics, supporting their clinical differentiation. These findings highlight the potential of sleep dynamics as a non-invasive diagnostic tool and may suggest differing therapeutic needs tailored to the unique sleep disruptions observed in these conditions. Our findings should not be directly generalized to males and older subjects, as our study population did not include them, necessitating further evaluations in these groups.

Chapter 7

Sleep-Stage Dynamics Predict Current Sleep-Disordered Breathing and Future Cardiovascular Risk

Abstract

Sleep-disordered breathing (SDB) is a major contributor to cardiovascular morbidity and disrupts both the macrostructure and dynamics of sleep stages (W, N1, N2, N3, REM). While specific alterations in sleep macrostructure, such as reduced durations of N3 and REM, have been linked to cardiovascular risk, the predictive value of sleep-stage dynamics remains unexplored. Using data from the prospective Sleep Heart Health Study, we applied a flexible forest-based modelling approach to a carefully selected cohort of 2579 subjects free from prior cardiovascular events and sleep-altering medications to minimize confounding. First, we demonstrate that a random forest classifier reliably identifies moderate-to-severe SDB (apnea–hypopnea index; AHI >15), achieving AUROC=76.1%, from sleep-stage architecture, dynamics, and common risk factors (demographics, BMI, smoking status) alone, without direct respiratory measurements. This highlights a dependency chain in which SDB correlates with altered sleep patterns that, in turn, encode cardiovascular risk. Second, a random survival forest robustly predicted future cardiovascular events (concordance-index=73.3%) over >10 years follow-up. Comparable results with and without including AHI as a predictor indicate that sleep patterns encode cardiovascular risk independently of direct SDB measurement. Partial dependence analyses revealed monotonic SDB risk profiles and predominantly U-shaped associations for cardiovascular risk, identifying ranges of total sleep time, wake after sleep onset, and REM/N3 continuity linked to minimal or elevated risk. Notably, rare transitions such as N3→N1 or REM→N3, even occurring once per night, emerged as sensitive markers of cardiovascular vulnerability, increasing risk by up to 10%. Our findings extend prior evidence on linear associations between sleep macrostructure and cardiovascular outcomes, revealing non-linear patterns and positioning sleep dynamics as promising non-invasive biomarkers for diagnostics and early risk stratification.

Keywords: *Cardiovascular risk, Sleep-disordered breathing, Sleep, Sleep-stage dynamics, Polysomnography, Machine learning, Explainable AI, Non-invasive biomarkers*

7.1 Introduction

Sleep-disordered breathing (SDB), particularly obstructive sleep apnea (OSA), is a well-established contributor to both cardiovascular morbidity [212]–[214] and mortality [2], [7], [215]. The pathophysiology of SDB involves intermittent hypoxia, intrathoracic pressure swings, and repeated arousals, which activate the sympathetic nervous system, trigger oxidative stress, and induce systemic inflammation and endothelial dysfunction—processes that collectively accelerate vascular remodelling and atherogenesis [5], [213], [214], [216]. In parallel, SDB promotes metabolic dysregulation through impaired insulin sensitivity and

altered adipokine signalling, contributing to obesity and further elevating cardiovascular risk [5], [213]. These mechanisms are reflected in many population-based studies. For example, moderate-to-severe SDB, defined by an apnea–hypopnea index (AHI) greater than 15, capturing the hourly rate of partial or complete breath arrests, was associated with a nearly 3-fold increased risk of incident ischemic stroke in men [212]. Similarly, SDB was prospectively linked with incident hypertension over four years, with over a 2-fold increased risk observed in those with AHI ≥ 5, independent of obesity and baseline blood pressure [4].

Beyond its contribution to cardiovascular morbidity, SDB is also associated with increased risks of all-cause and cardiovascular mortality, with nearly double the risk of stroke or death compared to individuals without the condition, even after adjusting for major confounders [215]. Severe SDB has also been linked to a 46% higher risk of all-cause mortality over long-term follow-up [2]. Importantly, SDB may alter the temporal distribution of deaths: sudden cardiac death in individuals with OSA is more likely to occur during the night, particularly between midnight and 6 AM, in contrast to the early morning peak seen in the general population [217]. A severe SDB has been associated with a 2.5-fold increased risk of sudden cardiac death [7].

Beyond respiratory disturbances, SDB leads to marked alterations in sleep macro-architecture, described by five sleep–wake states [8]: wakefulness (W), rapid eye movement (REM) sleep, and three non-REM (NREM) stages: light (N1, N2) and deep slow-wave (N3) sleep, each representing a distinct physiological state [13], [165]. Specifically, SDB is associated with reduced proportions of N3 and REM sleep, critical for physical and cognitive restoration, respectively, as well as increased time in wake after sleep onset (WASO), and frequent awakenings and micro-arousals that fragment sleep continuity [25], [218], [219]. These changes lead to lighter, less efficient sleep dominated by N1 and N2 stages, which contributes to daytime symptoms and heightened sympathetic activity—an aspect of autonomic imbalance linked to cardiovascular risk [25], [218]. These macro-structural abnormalities may serve as both markers and mediators of downstream cardiometabolic dysfunction.

In addition to altered stage composition, SDB disrupts the temporal continuity and organization of sleep stages—a feature referred to as sleep-stage dynamics [26], [106], [109], [115]. The dynamics describe how individuals transition between stages over time and may capture subtle signatures of sleep instability that are not evident in static macrostructure metrics such as stage proportions or durations. The dynamic patterns have been shown to reflect diverse physiological and pathological conditions beyond SDB [100]–[102], [104], [105], [107], [114], [116], [171]. Importantly, specific sleep-stage transitions may reflect underlying physiological needs: frequent transitions into deep slow-wave sleep (N3) signal elevated homeostatic sleep pressure and physical restoration processes [13], [220], while disrupted or shortened REM periods signal impaired cognitive and emotional recovery [221]–[223]. Prior work has shown that SDB-individuals exhibit irregular and less predictable sleep-stage transitions, accompanied by abnormal heart rate variability patterns [109]. Transition-based modelling approaches have revealed elevated transition entropy and reduced stage persistence, indicating a loss of normal sleep structure in SDB subjects [26]. Recent analyses have further demonstrated that these dynamic patterns vary systematically by age, gender, and apnea severity, with distinct fragmentation profiles across REM and NREM stages [106], [115].

Notably, disrupted sleep macro-architecture has been linked to increased cardiovascular risk and mortality. Reduced slow-wave sleep (N3) is associated with incident hypertension, possibly reflecting impaired nocturnal blood pressure regulation [37], [38], while lower REM sleep has been linked to higher all-cause and cardiovascular mortality [39]. Diminished delta wave activity during sleep, related to reduced N3, has also been associated with long-term cardiovascular outcomes [40]. Abnormal total sleep duration, as well as poor self-reported sleep quality, show associations with cardiovascular and all-cause mortality [41]–[44].

Although disrupted sleep macro-architecture has been repeatedly associated with cardiovascular outcomes and mortality [5], [37]–[44], [216], the prognostic relevance of sleep-stage dynamics—able to capture detailed physiological signatures—remains unexplored. Since SDB affects both cardiovascular outcomes [2], [4], [7], [212]–[215] and sleep-stage dynamics [26], [106], [109], [115], these two domains are statistically linked, suggesting that sleep

dynamics may encode predictive signals relevant for cardiovascular risk modelling. Furthermore, dynamic sleep patterns have been associated with a range of other conditions, including insomnia, chronic fatigue syndrome, pain syndromes (fibromyalgia), sleep bruxism, and neurocognitive impairment [100]–[102], [104], [105], [107], [114], [116], [171], which may also contribute to cardiovascular vulnerability. This interplay highlights the potential of sleep dynamics as integrative, non-invasive digital markers for diagnostics and long-term cardiovascular risk assessment.

Study contributions and research question. We present the first investigation to evaluate whether and eventually how sleep-stage dynamics, alongside conventional sleep macrostructure metrics and common risk factors (demographics, BMI, smoking status), carry prognostic value for long-term cardiovascular outcomes. Using data from the prospective, longitudinal, community-based Sleep Heart Health Study (SHHS) [182], we define a primary analysis cohort consisting of individuals without prior cardiovascular events and free from medications altering sleep architecture or cardiovascular physiology (e.g., antidepressants, beta-blockers, diuretics, aspirin). This design reduces confounding and enhances the generalizability of our findings to broader populations. For modelling, we adopt forest-based approaches [224]–[227], which offer a flexible non-parametric framework robust to overfitting, variable interactions, multicollinearity, and non-linear effects. We consider four groups of predictors, capturing static and dynamic properties of sleep and relevant risk factors: (i) *percentages of sleep-stage transitions* (e.g., W $\rightarrow$ N1) relative to the time after sleep onset characterizing sleep-stage dynamics; (ii) *conventional sleep metrics* (e.g., total sleep time [TST], WASO); (iii) *demographics* (age, gender); and known (iv) *risk factors* including body mass index (BMI) and smoking status.

Our study presents four key contributions:

1. **Identification of current SDB status:** We apply a Random Forest (RF) classifier to identify individuals with moderate-to-severe SDB (AHI > 15) and demonstrate that characteristic changes in sleep architecture and dynamics carry diagnostic information about underlying respiratory disturbance, and hence, a possible link to long-term cardiovascular health.

2. **Prediction of long-term cardiovascular risk:** We use a Random Survival Forest (RSF) to estimate long-term cardiovascular risk. By comparing models with and without an AHI predictor characterizing the SDB severity, we assess whether the four sets of predictors alone can encode the prognostic information attributable to SDB.

3. **Validation and generalizability:** We assess the RF and RSF models through cross-validation within the primary study cohort and across additional SHHS subgroups, including baseline or follow-up polysomnography (PSG) data from individuals with prior cardiovascular events or medication use. A further validation is performed using the external *Bern Sleep-Wake Registry* (BSWR). In BSWR, SDB predictions are directly evaluated, while cardiovascular risk estimates are regressed against various sleep disorders and other comorbidities.

4. **Interpretability and novel markers via partial effects:** We use partial dependence plots from the R(S)F models to assess how individual predictors influence both SDB and cardiovascular risk, revealing non-linear effects, candidate thresholds for clinical risk-stratification, and suggesting novel diagnostic and prognostic markers.

7.2 Materials and Methods

7.2.1 Data sets

Sleep Heart Health Study (SHHS)

The SHHS is a multi-center, prospective cohort study designed to investigate the cardiovascular consequences of sleep-disordered breathing (SDB) in middle-aged and older

adults [182]. The SHHS recruited participants from existing population-based cardiovascular and respiratory cohorts across the United States between 1995 and 1998. The baseline exam (SHHS1) was conducted on 6441 subjects and included in-home overnight polysomnography (PSG), comprehensive medical questionnaires, and cardiovascular assessments. A subset of 3295 participants underwent a second PSG study (SHHS2) about 5–8 years later as part of a follow-up evaluation. Participants have been longitudinally followed for over a decade to track major cardiovascular events and mortality. All PSG recordings in SHHS1 and SHHS2 were conducted without positive airway pressure (PAP) therapy, allowing participants to be considered untreated at the time of measurement. The SHHS data set comprises PSG biosignal data with sleep scoring annotations, detailed medication information, demographic and anthropometric measures, and details on the timing and type of cardiovascular outcomes. This study design enables the investigation of both cross-sectional and longitudinal relationships between sleep architecture and cardiovascular risk in a community-based population. In total, data from the 5839 participants who consented to share their information are available for research purposes.

Cardiovascular events: The outcome of interest was defined as the first occurrence of any major cardiovascular event recorded in SHHS and following the PSG assessment, including both clinical diagnoses and surgical interventions: angina, angioplasty, coronary artery bypass graft (CABG), congestive heart failure (CHF), myocardial infarction (MI), myocardial infarction procedure (MIP), percutaneous transluminal coronary angioplasty (PTCA), revascularization procedures, coronary stenting, or stroke. These events represent a mixture of atherosclerotic disease manifestations and interventional procedures commonly conducted in high-risk individuals. For survival analyses, we defined the event time as the number of days from the PSG recording (SHHS1 or SHHS2) to the first occurrence of any listed event. If no event occurred, the number of days to the most recent follow-up contact or recorded death since the PSG study was used as the censoring time. In addition, we identified whether individuals had experienced any of these cardiovascular events prior to the PSG study and used this information for stratification and assessments of the models' generalizability.

Medication-related confounders: To account for potential pharmacological confounding, we created a binary indicator variable at both baseline (SHHS1) and follow-up (SHHS2) identifying subjects who were taking medications known or suspected to alter sleep-stage composition or cardiovascular risk. Based on clinical expertise, we flagged use of medications from several categories listed in SHHS metadata: psychiatric agents (e.g., tricyclic antidepressants, monoamine oxidase inhibitors, other antidepressants, antipsychotics, benzodiazepines), neurological agents (e.g., dopaminergic medications for Parkinson's disease, cholinesterase inhibitors for Alzheimer's disease), and selected cardiovascular drugs (e.g., beta-blockers, alpha-blockers, ACE inhibitors with diuretics, vasodilators, loop and thiazide diuretics). Aspirin was also included due to its high use and reported influence on slow-wave sleep. Medication status of subjects was extracted using SHHS drug codes, and the resulting indicator variable was used to support stratification and generalizability assessments across subgroups with and without pharmacological confounding.

Cohort stratification and notation: We analyzed a total of 8442 PSG recordings, comprising 5791 baseline recordings from SHHS1 and 2651 follow-up recordings from SHHS2. Each of these PSGs of unique individuals was successfully linked to available clinical and demographic metadata, including medications, cardiovascular event histories, event dates, and censoring information. To support subgroup analyses and generalizability tests, we organized the data according to three key attributes: (i) *study wave* (SHHS1 or SHHS2); (ii) *prior cardiovascular events* at the time of PSG recording (**E = 1** if any event occurred before PSG, **E = 0** otherwise); and (iii) *presence of medications* known to influence sleep or cardiovascular physiology (**M = 1** if such medications were reported, **M = 0** if not). The Supplementary Table C.4 presents the number of subjects/PSGs (N) in each stratum, along with the number of subjects who developed an event following the PSG study, and the distribution of subject ages and genders. Our modelling efforts focused primarily on baseline subjects with no previous events or confounding medications, SHHS1(E = 0, M = 0), enabling the most precise evaluation of how specific sleep or demographic patterns associate with current SDB and the development of future cardiovascular events, not confounded by medication intake or prior events. The remaining subsets of data were used for validation and robustness assessments.

Bern Sleep-Wake Registery (BSWR)

To assess external validity and generalizability, we exploited the Bern Sleep-Wake Registry (BSWR) from Inselspital, University Hospital Bern. The BSWR contains over two decades of clinical polysomnography (PSG) data, starting from 2000. Most individuals in the BSWR suffer from one or more sleep disorders, with annotations including demographic information, clinical diagnoses, and relevant comorbidities beyond sleep-related conditions. For this study, we excluded daytime PSG recordings, studies shorter than 3 hours, instances where patients failed to fall asleep, and recordings involving positive airway pressure (PAP) therapy. Our final data set included 3702 PSG recordings from 3417 unique individuals aged 0–91 years (62.8% males), with a conclusive sleep diagnosis, complete demographic data (age, gender), and a calculated apnea–hypopnea index (AHI) to quantify the severity of sleep-disordered breathing (SDB). We primarily used the BSWR for external validation of moderate-to-severe SDB detection (AHI $>$ 15). As BSWR currently lacks harmonized and matched time-to-event histories of cardiovascular outcomes, it was not feasible to use it for direct validation of long-term cardiovascular risk prediction. Instead, we leveraged the rich clinical annotations within the BSWR to evaluate associations between the model-predicted cardiovascular risk and specific sleep diagnoses, as well as relevant non-sleep comorbidities (e.g., prior stroke, diabetes). This strategy enabled both intuitive clinical validation and an indirect quantification of how individual clinical conditions relate to predicted cardiovascular risk. The Supplementary Table C.2 presents the occurrence of different clinical conditions, including conclusive sleep disorders and non-sleep comorbidities, across BSWR subjects, and their stratification by sleep-disordered breathing status (AHI$\leq$15 vs AHI>15).

Data Preprocessing

All PSG recordings from both the SHHS and BSWR cohorts were scored into five standard sleep stages: Wake, N1, N2, N3, and REM, according to AASM guidelines [8]. Older recordings originally scored using the Rechtschaffen and Kales (R&K) guidelines were harmonized by merging the N3 and N4 stages into a single AASM-compliant N3 stage. The apnea–hypopnea index (AHI) was computed using the *recommended* AASM definition (v2.2, 2015) and used to derive binary sleep-disordered breathing (SDB) labels, with moderate-to-severe SDB defined as AHI > 15 and no-to-mild SDB defined as AHI$\leq$15. Gender was encoded as a binary male indicator (1 = male, 0 = otherwise); in all cases, the non-male category corresponded to participants self-identifying as female. Smoking status was encoded as a categorical variable with four levels: *current*, *ex*, *never*, or *not-available* (NA). Established sleep macrostructure features—such as total sleep time (TST), Wake After Sleep Onset (WASO), and stage-specific latencies—were computed directly from the hypnograms. Sleep dynamics were captured, following our prior work [115], as a 5$\times$5 matrix of sleep-stage transition proportions $\mathbf{P}$, where each entry $p_{i,j}$ denotes the percentage of all epochs, relative to the time after sleep onset, during which a transition from stage i to stage j occurred. For cardiovascular risk modelling, survival objects were constructed using the time from the PSG study (either baseline SHHS1 or the follow-up SHHS2) to the first joint cardiovascular event or censoring at the last contact.

7.2.2 Prediction, Validation, and Effect Quantification using Random (Survival) Forests

Modelling approach and predictors. To detect moderate-to-severe SDB (AHI>15), we employed a binary *Random Forest* (RF) classifier. For long-term cardiovascular risk prediction, we used *Random Survival Forests* (RSF), a non-parametric extension of RF for right-censored time-to-event data using the log-rank test as a splitting criterion [227]. Both RF and RSF are ensemble-based approaches, known for their robustness to overfitting, ability to capture non-linear relationships, and resilience to multicollinearity and high-dimensional settings (cf. [224]–[226]), making them particularly advantageous when using many correlated predictors such as the 25 sleep-stage transition proportions ($\mathbf{P}$), where functional dependencies exist as all sum up to 100%. Each RF or RSF model was trained using the following predictor sets: *sleep dynamics* captured by 25 transition proportions, *conventional sleep metrics* not

encoded in dynamics features: TST, WASO, and (sleep, N3, REM)-latencies, and *demographic or lifestyle variables:* age, gender, BMI, and smoking status. Additionally, for RSF, we experimented by adding AHI as an additional predictor to compare performance in cardiovascular risk prediction with and without the inclusion of direct SDB measurement. For both RF and RSF, we used default values of hyperparameters provided by the `randomForestSRC` R package (v3.1.1), which have been shown to perform reliably [228].

Validation was carried out in three tiers. *First*, we used 5-fold cross-validation (CV) on the primary SHHS-baseline cohort of subjects without prior cardiovascular events or medication use, SHHS1(E = 0, M = 0), with approximately equal-sized folds created using fast anticlustering [229], balancing the distribution of demographics and outcomes (age, gender, AHI, SDB- and cardiovascular-events prevalence). *Second*, models were tested on the remaining cohorts from SHHS1 and SHHS2 (cf. Supplementary Table C.4), with additional control for subjects used during the models' training. This assessed performance and discriminative power in the same subjects but several years later, i.e., by using SHHS2(E = 0, M = 0), and also in out-of-domain subjects with medications or prior events, which typically have higher rates of SDB prevalence, cardiovascular events, and different distributions of demographics (cf. Supplementary Tables C.5-C.11). *Third*, external generalization was tested on a completely out-of-domain BSWR clinical data set, which, unlike SHHS, primarily contains a symptomatic clinical population suffering from multiple sleep disorders and non-sleep comorbidities. Selection of subjects with evaluated AHI enabled direct generalizability assessment of RF model in BSWR. The log-odds-transformed RSF-predicted cardiovascular risk in BSWR was regressed against different clinical conditions present in BSWR, while controlling for age, gender, BMI, and AHI, allowing for assessments of possible links between various conditions and an elevated cardiovascular risk.

Performance metrics for RF classification included the Area Under the Receiver Operating Characteristic curve (AUROC), Pearson's correlation between the predicted probability of moderate-to-severe SDB and the observed AHI, as well as accuracy, sensitivity, specificity, and precision. For RSF, model performance was evaluated using Harrell's concordance index (C-index), the Integrated Brier Score (IBS), and time-dependent AUROC (tdAUROC), each assessing the model's ability to rank individuals according to their actual risk. Discriminative ability of RSF was further assessed via two-sample t-tests comparing the mean predicted *mortality* scores—interpreted as individual risk estimates scaled to event frequency [228]—between those who did and did not experience a future event. Lastly, log-rank tests were used to compare event incidence between high- and low-risk groups, stratified by the median predicted mortality.

Towards novel markers through partial effects. Beyond evaluating performance of R(S)F models, we assessed their *partial effects* quantifying the isolated contribution of each predictor to model output. Specifically, we used the `partial.rfsrc` function from the `randomForestSRC` package [228], which quantifies changes in expected model prediction when one predictor is varied over its domain while all other features are held fixed and averaged over their empirical joint distribution. For RF, this yielded the partial effect of each variable on the probability of present moderate-to-severe SDB; for RSF, it revealed how each predictor influences the long-term (10 years) cardiovascular risk. These partial dependence functions provide interpretable insights into potential non-linearities, plateaus, or U-shaped effects, enabling models' explanation and importantly, supporting clinical interpretation of the relationship between specific sleep (dynamics and macrostructure) parameters and cardiopulmonary outcomes. Based on this, partial effects may serve as future diagnostic or risk stratification markers.

7.3 Results

The *Results* section is organized following our main objectives: *(i) Descriptive statistics*, summarizing demographics, SDB prevalence, and clinically established sleep markers, stratified by the occurrence of future cardiovascular events in the primary study cohort used for model development; *(ii) Performance of the RF classifier* for identifying current moderate-to-severe SDB (AHI >15); *(iii) Performance of RSF for predicting cardiovascular risk*, assessed both

with and without inclusion of the AHI predictor. Sections (ii) and (iii) also report the partial effects of individual predictors to investigate their risk associations, and evaluate model performance across SHHS test subgroups and the external BSWR data set.

7.3.1 Descriptive statistics

Table 7.1 summarizes demographic characteristics, SDB prevalence and severity, and sleep macro-structure metrics for the primary SHHS1(E = 0, M = 0) cohort, which excludes individuals with prior cardiovascular events (E = 1) or sleep-altering medications (M = 1). Statistics are presented for the entire cohort (*overall*) and stratified by the occurrence of a cardiovascular event during the follow-up period of up to 15 years. Cardiovascular events refer to a pooled composite outcome, including diagnoses such as stroke, myocardial infarction, heart failure, or surgical procedures such as coronary revascularisation (cf. *Data sets*).

Out of the 2579 subjects, 326 experienced at least one cardiovascular event during follow-up. When compared to subjects who did not develop the event, these individuals were significantly older at baseline (mean age 68.6 vs. 58.1 years), more likely to be male (55.5% vs. 44.4%), and had a different composition of smoking status profiles (e.g., 14.1% vs. 10.4% current smokers, and 43.9% vs. 49.5% never smokers). BMI did not differ between the compared groups, potentially due to group differences in demographics and their interactions with smoking status. These demographic and lifestyle differences are consistent with established cardiovascular risk factors [4], [214], [230]–[233].

Subjects who developed events had higher mean AHI values (18.3 vs. 15.7 events/hour) and greater prevalence of moderate-to-severe SDB (47.5% vs. 37.5%), reflecting the link between SDB and cardiovascular risk [2], [5], [7], [212]–[216]. However, since AHI increases with age, this association may also reflect age distribution differences [230], [231].

Sleep macrostructure also differed: those who developed events had shorter TST and longer WASO, resulting in lower sleep efficiency (SE = $\frac{\text{TST}}{\text{SL + TST + WASO}}$). They also showed altered sleep-stage composition, with higher W%, and reduced N3 and REM%—stages critical for physical and cognitive restoration. These changes likely reflect both elevated cardiovascular risk and age- or SDB-related alterations of sleep structure.

These statistics provide a descriptive overview of group-level trends and are not intended as clinically conclusive findings. Rather, they highlight the need for a more flexible modelling approach that can simultaneously account for multiple confounding variables, capture their interactions, and accommodate potential non-linear effects. Accordingly, in the following sections, we employ Random Forest and Random Survival Forest models to numerically isolate the contributions of key risk factors—such as age, smoking, BMI, and SDB severity—to cardiovascular outcomes and sleep-related metrics.

Table 7.1: Descriptive characteristics of SHHS1 (E = 0, M = 0) cohort stratified by cardiovascular event status.

Variable	Overall	Event-free	Event developed	p-value
N	2579	2253	326	
Age	59.43 (11.16)	58.11 (10.59)	68.57 (10.73)	<0.001
Gender (Male)*	1182 (45.8)	1001 (44.4)	181 (55.5)	<0.001
Smoking*				0.041
Current	281 (10.9)	235 (10.4)	46 (14.1)	
Ex	1010 (39.2)	874 (38.8)	136 (41.7)	
Never	1259 (48.8)	1116 (49.5)	143 (43.9)	
NA	29 (1.1)	28 (1.2)	1 (0.3)	
BMI	27.73 (4.92)	27.66 (4.96)	28.21 (4.63)	0.062
AHI	16.00 (14.76)	15.66 (14.59)	18.34 (15.71)	0.002
SDB (AHI>15)*	999 (38.7)	844 (37.5)	155 (47.5)	0.001
SDB category*				0.012
Mixed	429 (16.6)	362 (16.1)	67 (20.6)	
NREM-dominant	78 (3.0)	68 (3.0)	10 (3.1)	
REM-dominant	369 (14.3)	310 (13.8)	59 (18.1)	
AHI≤15	1580 (61.3)	1409 (62.5)	171 (52.5)	
NA	123 (4.8)	104 (4.6)	19 (5.8)	
TST [mins]	365.10 (63.93)	366.36 (63.09)	356.33 (68.90)	0.008
WASO [mins]	90.14 (54.66)	88.46 (53.80)	101.74 (59.09)	<0.001
SE [%]	72.29 (12.01)	72.53 (11.79)	70.64 (13.34)	0.008
SL [mins]	50.63 (43.05)	51.11 (43.18)	47.31 (42.00)	0.136
REML [mins]	104.81 (140.70)	104.30 (136.67)	108.35 (166.06)	0.627
DL [mins]	60.90 (185.96)	58.67 (182.37)	76.32 (208.77)	0.109
W [%]	19.65 (11.57)	19.28 (11.31)	22.19 (12.95)	<0.001
N1 [%]	4.02 (2.84)	3.99 (2.81)	4.20 (3.05)	0.200
N2 [%]	45.34 (11.43)	45.28 (11.26)	45.78 (12.52)	0.455
N3 [%]	14.70 (9.50)	14.95 (9.49)	12.93 (9.35)	<0.001
REM [%]	16.30 (6.13)	16.50 (6.10)	14.89 (6.14)	<0.001

Notes: Continuous variables are reported as mean (SD) and compared using Welch's two-sample *t*-test. Categorical variables, denoted by superscript *, are reported as counts (percentages) and compared using the chi-squared test. When expected cell counts were less than 5, Fisher's exact test was used instead.

7.3.2 Identification of current SDB status

Predictors and training of RF

To identify the presence of SDB, we trained an RF classifier using a total of 34 predictors: 25 sleep-stage transition proportions capturing sleep dynamics; 5 conventional sleep macrostructure metrics not encoded in dynamics (TST, WASO, and latencies to sleep onset, N3, and REM); and 4 commonly recognized risk factors accounting for demographic and lifestyle variation (age, gender, BMI, and smoking status). We did not include SE, as it is a function of TST, WASO, and SL. The binary outcome label indicated moderate-to-severe SDB (positive class, AHI>15) versus no-or-mild SDB (negative class, AHI≤15), supported by evidence linking AHI>15 to substantially elevated risks of stroke, cardiovascular morbidity, and mortality [2], [7], [212].

Given the imbalance between 38.7% of positive and 61.3% of negative cases, we used a quantile-based RF implemented in `randomForestSRC` R package [228], [234], effectively mitigating class dominance. To quantify the uncertainty in performance metrics (e.g., AUROC, accuracy), we performed repeated training using 5-fold cross-validation on anticluster-based data splits (see *Methods*) of the primary cohort SHHS1(E = 0, M = 0). For generalization assessments on the remaining SHHS subsets and BSWR, and the interpretation of partial effects, a final RF model was estimated on the complete set of 2579 subjects.

All RF models used default hyperparameters: a minimum terminal node size of 1, 10 candidate split points per variable, an AUROC-based splitting criterion, and the square root

of the number of considered predictors as the number of variables to try to split each node. Each forest consisted of 1000 trees, and missing input values were imputed using the built-in out-of-bag terminal node imputation algorithm [227], [234].

Performance and generalization of RF

The first part of Table 7.2 reports the mean (standard deviation) performance metrics in the primary cohort SHHS1(E=0, M=0), assessed using cross-validation. Across five anticluster splits, the data set contained approximately 316 controls (AHI$\leq$15) and 200 positive cases (AHI$>$15) on average. The RF classifier demonstrated strong discriminative ability, achieving an AUROC of 76.1 (2.2)% for distinguishing moderate-to-severe SDB cases from healthy-to-mild controls based on sleep parameters (i.e., transition %'s, macrostructure) combined with other variables (demographics, BMI, smoking status) alone. The strong performance was further supported by a strong positive correlation of 0.45 (0.05) between the predicted probabilities of the positive class and the subject's actual AHI values. The RF achieved an accuracy of 70.0 (1.6)%, sensitivity of 46.8 (4.6)%, specificity of 84.7 (1.1)%, and precision of 65.9 (1.9)%. Considering that SDB is typically quantified using respiratory and oxygen desaturation signals, the obtained performance metrics indicate a favourable balance between case detection and a low false positive rate.

Table 7.2 also presents performance metrics for different SDB phenotypes: *REM-dominant* (AHI$>$15 in REM sleep only; $\sim$74 cases), *NREM-dominant* (AHI$>$15 in non-REM sleep only; $\sim$16 cases), and *Mixed* (AHI$>$15 in both REM and NREM sleep; $\sim$86 cases). Although the original RF was optimized for overall SDB discrimination (i.e., AHI$>$15 vs. AHI$\leq$15), the AUROC remained high across these subgroups: approximately 74% for REM- and NREM-dominant SDB and 79.4% for *Mixed* cases. Other performance metrics were similarly robust, with a drop in precision for NREM-dominant SDB due to the small number of cases.

Next, Table 7.2 provides 95% confidence intervals (CIs) for AUROC and correlation coefficients, along with performance metrics for additional test subsets from the SHHS1 baseline study (including participants on medications or with prior cardiovascular events) and the SHHS2 follow-up. Stratification further considered whether PSG recordings came from previously unseen (out-of-domain) individuals, not included in RF training, highlighted by the † superscript. The results indicate that SDB detection remained robust even for SHHS1 participants with medications or prior events, achieving AUROC values >71% in all cases. Similar performance was observed across all subsets in the follow-up study: for the SHHS2 participants without medications or prior events used in RF training, SHHS2(E = 0, M = 0), AUROC reached 77.3%, and in unseen test follow-up medication- and event-free participants, SHHS2†(E = 0, M = 0), performance further improved to 80.6%, underscoring the model's strong generalizability and robustness to domain shifts, as the follow-up participants are naturally older from their baseline assessment. The performance on all test strata is particularly strong, given that post-event or medicated individuals exhibit altered sleep patterns [235]–[238], and these cohorts are on average older (cf. Supplementary Tables C.5-C.11).

Table 7.2: Random Forest identification of moderate-to-severe sleep-disordered breathing across SHHS and BSWR test datasets.

Dataset	AHI>15 (N)	AHI≤15 (N)	AUROC	$\rho(P_{\text{AHI>15}}, \text{AHI})$	Accuracy	Sensitivity	Specificity	Precision
SHHS1$^{\text{CV}}$(E = 0, M = 0)	199.8 (1.1)	316 (1)	76.1 (2.2)	0.45 (0.05)	70.0 (1.6)	46.8 (4.6)	84.7 (1.1)	65.9 (1.9)
REM-dominant	73.8 (1.1)	316 (1)	74.1 (2.9)	0.45 (0.05)	76.9 (1.2)	43.0 (5.9)	84.7 (1.1)	39.6 (3.3)
NREM-dominant	15.6 (0.9)	316 (1)	74.9 (5.5)	0.37 (0.06)	83.1 (0.8)	50.0 (8.0)	84.7 (1.1)	13.9 (1.3)
Mixed	85.8 (0.8)	316 (1)	79.4 (3.2)	0.47 (0.07)	77.8 (1.3)	52.4 (7.1)	84.7 (1.2)	48.1 (3.0)
SHHS1†(E = 0, M = 1)	1179	1349	74.1 (72.2, 76.0)	0.43 (0.40, 0.46)	68.0	55.7	78.7	69.5
SHHS1†(E = 1, M = 0)	67	45	73.6 (63.9, 83.4)	0.47 (0.31, 0.60)	59.8	53.7	68.9	72.0
SHHS1†(E = 1, M = 1)	323	249	71.1 (66.9, 75.3)	0.44 (0.37, 0.50)	65.4	60.7	71.5	73.4
SHHS2†(E = 0, M = 0)	70	107	80.6 (73.8, 87.4)	0.56 (0.44, 0.65)	74.0	64.3	80.4	68.2
SHHS2†(E = 0, M = 1)	443	488	69.5 (66.1, 72.8)	0.42 (0.36, 0.47)	65.5	61.9	68.9	64.3
SHHS2†(E = 1, M = 0)	26	9	72.2 (53.6, 90.8)	0.48 (0.17, 0.70)	60.0	61.5	55.6	80.0
SHHS2†(E = 1, M = 1)	189	115	75.2 (69.6, 80.8)	0.47 (0.38, 0.56)	69.1	70.4	67.0	77.8
SHHS2 (E = 0, M = 0)	230	404	77.3 (73.5, 81.0)	0.53 (0.47, 0.59)	71.6	54.3	81.4	62.5
SHHS2 (E = 0, M = 1)	237	316	78.8 (75.1, 82.5)	0.49 (0.43, 0.55)	70.5	62.4	76.6	66.7
SHHS2 (E = 1, M = 0)	1	1	- (-, -)	1.0 (-, -)	50.0	100.0	0.0	50.0
SHHS2 (E = 1, M = 1)	5	10	68.0 (28.2, 100)	0.08 (-0.45, 0.57)	66.7	60.0	70.0	50.0
BSWR†	1602	2100	76.0 (74.5, 77.5)	0.49 (0.47, 0.52)	67.4	42.8	86.1	70.1

Notes: Evaluations included subjects with and without previous cardiovascular events (E = 0 and E = 1) or medications (M = 0 and M = 1), from baseline (SHHS1) and follow-up (SHHS2) Sleep Heart Health Study, and the Bern Sleep-Wake Registry (BSWR). Cross-validation assessment in the primary study cohort SHHS1 (E = 0, M = 0) is presented as mean (standard deviation) and highlighted by bold font and $^{\text{CV}}$ superscript. Datasets containing out-of-domain subjects are highlighted by †superscript. Results are based on the Random Forest classifier estimated on SHHS1(E = 0, M = 0). Table presents numbers (N) of cases (AHI>15) and controls (AHI≤15), 95% confidence intervals for Area Under the Receiver Operating Characteristic (AUROC) and Pearson's correlation coefficient of predicted probability of positive class ($P_{\text{AHI>15}}$) with actual AHI value, and the achieved performance metrics (%).

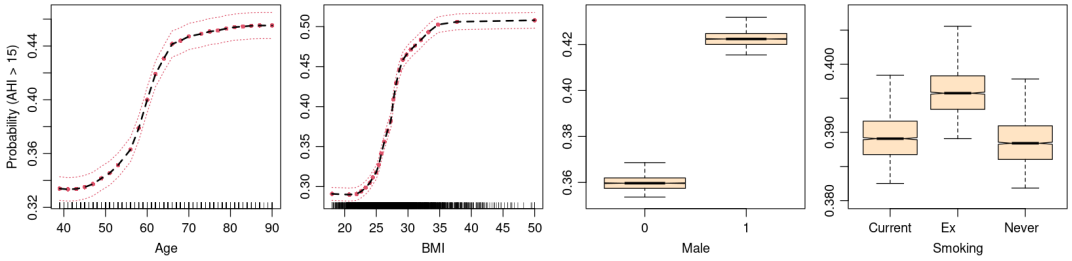
The RF maintained robust performance in the out-of-domain BSWR (Supplementary Tables C.1-C.2), which primarily includes patients with diverse sleep disorders rather than a general population sample, as in SHHS. Despite the altered sleep patterns expected in BSWR, the RF achieved an AUROC of 76.0%, a correlation of 0.49 between predicted probability and AHI, an accuracy of 67.4%, sensitivity of 42.8%, specificity of 86.1%, and a precision of 70.1%. These results highlight strong discriminative ability and high precision, indicating that individuals with high predicted probabilities are frequently true SDB cases.

SDB risk-profiles via partial effects of RF

As described in *Methods*, the trained RF enables quantification of partial effects for individual predictors X_j. These effects mathematically represent the change in the RF output (i.e., the predicted probability of moderate-to-severe SDB) when iterating over all (or, for computational efficiency, a subset of) observed values of X_j, substituting these and averaging predictions across all N observations in the data set while holding other variables fixed. As N predicted outcomes are obtained for each considered value of X_j, the expected partial effect is estimated as the mean, and uncertainty is expressed using quantile-based confidence intervals computed around this estimate. Plotting partial effects and their associated intervals yields a "risk profile" that shows how variations in X_j relate to the RF-predicted SDB-probability.

Figure 7.1 illustrates the partial effects of demographic (age, gender) and lifestyle (BMI, smoking) variables on the probability of moderate-to-severe SDB (AHI≥15). Both age and BMI exhibit sigmoidal profiles, reflecting an accelerated risk increase within specific ranges before reaching a plateau. For age, the predicted SDB probability rises sharply from 33% to 45% between 45 and 70 years. A similar, but even steeper, trend is observed for BMI: the risk increases from about 30% at BMI ≤25 to >50% for BMI>35. Partial effects further indicate a 6% higher predicted risk in males (36% in females vs. 42% in males). These associations are consistent with established age, obesity, and gender differences in SDB prevalence[230], [231], [239]. Interestingly, ex-smokers show a 1% higher predicted risk compared to both current and never-smokers, potentially reflecting post-cessation weight gain[232], [233] or underlying health issues prompting cessation. It is important to note that these partial effects

Figure 7.1: Partial effects and their 95% CIs for the risk of moderate-to-severe sleep-disordered breathing (AHI>15) for the age in years, Body Mass Index (BMI), gender (0 = female, 1 = male), and smoking status.



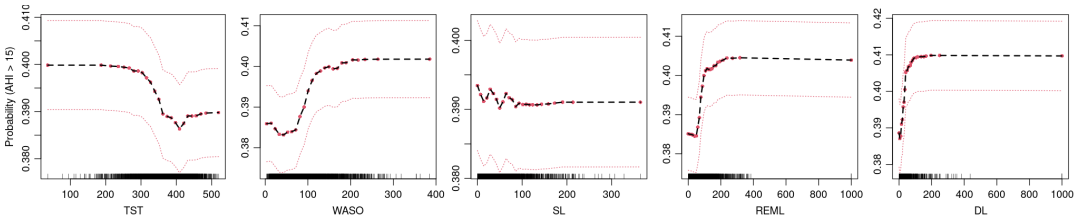
Notes: Data points are shown as ticks on the x-axis.

are contingent upon the population characteristics of the RF training cohort. In particular, the SHHS design oversampled individuals with snoring to increase statistical power for long-term cardiovascular outcomes [182]. Consequently, baseline SDB risk estimates (e.g., for individuals in their early 40s) may be inflated relative to the general population.

Figure 7.2 shows the partial effects of sleep macro-architecture markers used as RF inputs. The SDB risk increases markedly when TST falls below 300 minutes (5 hours) and when WASO exceeds 100 minutes, reflecting reduced sleep efficiency and prolonged wakefulness, likely due to apnea-related arousals. In contrast, the sleep-onset latency (SL) shows no clear association with SDB risk, whereas delays in entering REM sleep (REML) and deep sleep (DL) beyond 100 minutes are associated with an approximate 2% increase in risk.

Figure 7.3 details partial effects for sleep-stage transition proportions $p_{i,j}$ (cf. *Methods*), computed relative to the total number of sleep-stages from sleep onset to the end of the PSG

Figure 7.2: Partial effects and their 95% CIs for the risk of moderate-to-severe sleep-disordered breathing (AHI>15) for the minutes of Total Sleep Time (TST), Wake After Sleep Onset (WASO), Sleep Latency (SL), REM Latency (REM), and Deep-sleep Latency (DL).



Notes: Data points are shown as ticks on the x-axis.

recording. As shown in Table 7.1, the average after-onset PSG-duration (sleep period time) of 455.2 minutes (TST + WASO = 365.10 + 90.14) corresponds to approximately 910 epochs, such that a 1% (= 0.01) change in transition proportion corresponds to roughly 9 transitions per PSG. For interpretability, we focus on transitions associated with ≥2% change in predicted risk, supported by non-overlapping deviations in 95% CI profiles of their partial effects.

- *W-transitions*: The SDB risk increases by ∼2% when $p_{W,N1} > 0.04$ (∼36 transitions), reflecting frequent awakenings followed by light sleep in SDB subjects. Similarly, $p_{W,N2} > 0.02$ (∼18 transitions) is associated with >2% higher risk, suggesting that SDB subjects experience heightened homeostatic sleep pressure and atypically bypass the intermediate N1 stage after awakenings. Notably, $p_{W,REM} > 0.08$ corresponds to a dramatic risk increase (>5%), indicating that direct transitions from wake to REM, possibly related to higher restorative pressure or REM sleep fragmentation, are highly sensitive markers of SDB.

- *N1-transitions*: The SDB-risk increases by >6% when $p_{N1,W} > 0.01$ (∼9 transitions), suggesting frequent arousals of SDB subjects from lightest sleep. Notably, $p_{N1,REM} > 0.005$ (∼4 transitions) is linked to a ∼2% higher risk, possibly reflecting atypical transitions to REM from N1 in SDB patients, similar to those from W.

- *N2-transitions*: Higher proportions of $p_{N2,W} > 0.03$ (∼27 transitions) correspond to a ∼8% SDB-risk increase, indicative of frequent awakenings from N2, likely linked to disruptions driven by experienced apnea events.

- *N3-transitions*: A SDB-risk increase ∼3% is observed when $p_{N3,N3} < 0.1$, corresponding to ∼90 uninterrupted epochs of N3, suggesting that SDB subjects are often unable to achieve more than 45 minutes of continuous deep sleep over the entire night.

- *REM-transitions*: The SDB-risk rises by ∼3% when $p_{REM,W} > 0.015$ (∼13 transitions), highlighting increased REM-awakenings in SDB. In contrast, higher persistence within REM ($p_{REM,REM} > 0.2$, ∼180 epochs, = 1.5 hours) links to a lower risk, suggesting difficulties in retaining uninterrupted REM-sleep in SDB.

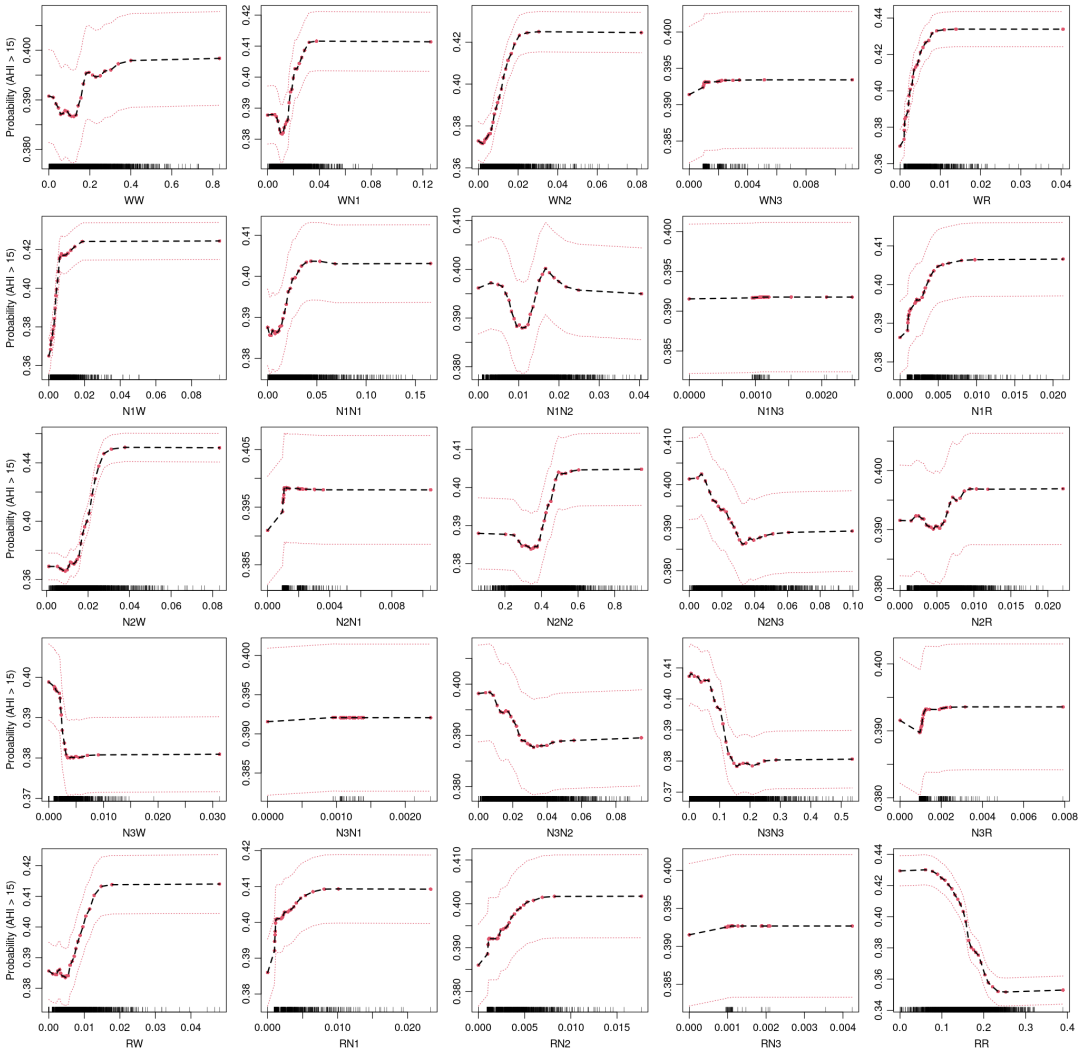
The estimated partial effects mechanistically explain RF decisions when identifying SBD. In addition, they characterize risk profiles, which may suggest thresholds that can be used as clinical instruments (markers) for diagnostics and risk stratification.

7.3.3 Prediction of long-term cardiovascular risk

Predictors and training of RSF

To quantify the risk of future cardiovascular events, we trained a RSF using the same 34 predictors as for RF, including 25 sleep-stage transition proportions, 5 sleep macro-structure metrics (TST, WASO, and latencies to sleep onset, N3, and REM), and 4 demographic and lifestyle variables (age, gender, BMI, and smoking status). To evaluate the added prognostic value of including a direct measure of SDB, we trained RSF models both with and without apnea–hypopnea index (AHI) as a predictor.

Figure 7.3: Partial effects and their 95% CIs for the risk of moderate-to-severe sleep-disordered breathing (AHI>15) for relative frequencies of individual transitions between sleep-stages (W, N1, N2, N3, REM = R).



Notes: Each subplot's x-axis label indicates the direction of the transition (e.g., WN1 corresponds to transitions from W to N1). Data points are shown as ticks on the x-axis.

As described in *Methods*, the RSF target was a survival outcome comprising an event indicator (1 if the participant experienced a cardiovascular event after the sleep study, 0 otherwise) and the time-to-event, defined as the number of days since the PSG study to either the first event or the most recent follow-up, if censored. The RSF aimed to map the links between survival outcomes and the included predictors, and to quantify their influence via partial dependence analysis subsequently.

The RSF was trained using `randomForestSRC` R package [228], [234] on the primary cohort SHHS1(E = 0, M = 0), to minimize confounding and maximize the generalizability to a general, medications- and prior-event-free population. To assess uncertainty in predictive performance and discrimination metrics (e.g., time-dependent AUROC [tdAUROC], Integrated Brier Score [IBS], Harrell's concordance index [C-index], log-rank test), we applied a 5-fold cross-validation with anticluster-based splits (*see Methods*). For generalization assessments on remaining SHHS subsets and BSWR, and interpretation of partial effects, a final RSF was trained on the complete set of 2579 subjects.

During training, default RSF hyperparameters were used: a minimum terminal node size of 15, 10 candidate split points per variable, C-index splitting criterion, and the square root of the number of predictors as the number of variables to try to split each node. Each RSF

comprised 1000 trees, and missing input values were imputed using the built-in out-of-bag terminal node imputation algorithm [227].

Performance and generalization of RSF

Table 7.3 and Supplementary Table C.12 summarize the performance and discrimination metrics of RSF trained while including and omitting the AHI as a predictor, respectively. Results are shown for the primary study cohort and for subjects across SHHS1 and SHHS2 subsets who were free of prior cardiovascular events at the time of PSG assessment. The first column of each table, labelled SHHS1^{CV}(E = 0, M = 0), reports cross-validation results within the primary cohort, expressed as a mean (standard deviation) for each metric. Each anticluster fold comprised, on average, about 516 subjects, of whom 65 developed a cardiovascular event during follow-up and 451 remained event-free. The RSF's ability to rank individuals by their actual cardiovascular risk was evaluated using three standard metrics: C-index, IBS, and tdAUROC. For C-index and tdAUROC, values of 100% indicate perfect discrimination, 50% random chance, and 0% inverted predictions. In contrast, lower IBS values indicate better calibration. For all metrics, cardiovascular risk was quantified via the RSF-predicted *mortality score*, representing an individual's standardized risk relative to a hypothetical population of subjects of matched characteristics [228], [234].

Table 7.3: Performance of the Random Survival Forest model including AHI predictor across SHHS and BSWR datasets of subjects with no previous cardiovascular events (E = 0).

Metric	SHHS1^{CV} (E = 0, M = 0)	SHHS1† (E = 0, M = 1)	SHHS2 (E = 0, M = 0)	SHHS2† (E = 0, M = 0)	SHHS2 (E = 0, M = 1)	SHHS2† (E = 0, M = 1)
Events (N)	65.2 (2.4)	567	43	19	64	137
Event-free (N)	450.6 (2.7)	1961	591	158	489	794
C-index	73 (2.5)	69.7	74.5	79.3	70.6	66.6
IBS	6.7 (0.4)	12	4.1	6.2	7	8.8
1-year tdAUROC	77.4 (12.4)	69.7	85	78.1	72.4	64.5
5-year tdAUROC	74.5 (6.6)	72.6	73.7	84.5	71.7	69.2
10-year tdAUROC	75.1 (2)	74.1	91.6	100	75.1	69.4
Mortality (E = 1)	20.8 (1.9)	23.4	25.2	35.6	29.5	31.5
Mortality (E = 0)	11.4 (0.5)	14.5	14.1	15.8	17.9	21.7
Mortality Diff.	9.5 (2)	8.8	11.1	19.8	11.6	9.8
Mortality Diff. CI-low	5.6 (1.5)	7.5	5.0	10	6.7	6.4
Mortality Diff. CI-high	13.3 (2.5)	10.2	17.3	29.6	16.6	13.3
p-value (t-test)	0.000024 (0.000034)	$< 10^{-6}$	0.0007	0.000422	0.000014	$< 10^{-6}$
Events (N), high-risk	49.6 (2.7)	403	35	16	49	95
Events (N), low-risk	15.6 (2.3)	164	8	3	15	42
χ^2	23 (6.1)	157.6	19.5	10.5	22.5	28.1
p-value (log-rank test)	0.000047 (0.000103)	$< 10^{-6}$	0.00001	0.001221	0.000002	$< 10^{-6}$

Notes: The ^{CV} superscript denotes performance obtained via 5-fold cross-validation (CV) on the in-domain event- and medication-free (E = M = 0) baseline cohort SHHS1(E = 0, M = 0). All other columns evaluate the performance of the final RSF model fitted to the entire baseline cohort, applied to potentially out-of-domain subjects (†) from either the baseline (SHHS1) or follow-up (SHHS2) studies, including subgroups taking medication (M = 1). For each scenario, the number of subjects with events and without events (event-free) is reported. Model performance is assessed using Harrell's Concordance Index (C-index), Integrated Brier Score (IBS), and the time-dependent Area Under the Receiver Operating Characteristic curve (tdAUROC) at 1, 5, and 10 years. Discriminatory ability is evaluated via two-sided t-tests comparing predicted mortality between event and non-event subjects, including 95% confidence intervals (CI) for the difference (Diff.). Additionally, log-rank tests with Chi-squared (χ^2) statistics compare event rates between high- and low-risk groups stratified by median predicted mortality. For the in-domain CV, mean (SD) of all metrics is reported.

Notably, a comparison of Table 7.3 and Supplementary Table C.12 reveals no measurable performance gain from including AHI. The performance metrics achieved with vs. without AHI predictor were statistically identical: C-index at 73.0 (2.5) vs. 73.3 (2.5), IBS at 6.7 (0.4) for both cases, and the (1, 5, 10)-year tdAUROC at [77.4 (12.4), 74.5 (6.6), 75.1 (2.0)] vs. [77.1 (12.2), 74.9 (6.2), 75.3 (2.3)], respectively. Considering a 2σ bound around the mean as an indicator of significant difference, there is not a single case in which the two models would differ. Comparable performance between the two RSFs is further confirmed by a discriminatory assessment using a t-test, which compares the mean difference in predicted mortality between subjects who developed an event and those who did not, and a log-rank test comparing high- and low-risk individuals identified by the median threshold on predicted mortality. For RSF with and without AHI, the cross-validation-averaged mean differences were 9.5 (95% CI: 5.6–13.3, p-val $\sim 2.4 \times 10^{-5}$) and 9.8 (95% CI: 5.8–13.7, p-val $\sim 1.3 \times 10^{-5}$), respectively, whereas the log-rank χ^2 statistics were 23 (p-val $\sim 4.7 \times 10^{-5}$) and 26 (p-val $\sim 5 \times 10^{-6}$), respectively. Finally, Supplementary Figures C.1 and C.13 illustrate the distribution of the primary study population concerning cardiovascular cases, survivors, and censoring, as well as the performance metrics (tdAUROC, IBS) for the RFS with and without AHI, respectively, yielding identical trends.

This robust finding suggests that, despite the well-known clinical evidence linking SDB to future cardiovascular events [2], [5], [7], [212]–[216], directly including AHI as a predictor does not provide any measurable improvement in model performance when common risk factors (demographics, BMI, smoking status), sleep macro-architecture and dynamics are already controlled for. This observation can be interpreted from two complementary perspectives. First, as shown in the previous experiment, the same set of predictors can effectively capture SDB-related patterns. It is therefore likely that correlations between SDB and cardiovascular risk are indirectly encoded in these features, allowing the RSF to leverage this information without requiring explicit measurement of AHI. Second, the onset of cardiovascular events may be influenced by broader alterations in sleep macrostructure and dynamics that extend beyond SDB-specific patterns. The RSF, being a flexible and highly capable modelling approach (1000 trees), is well-suited to detect such complex relationships. These could include subtle patterns reflecting the downstream effects of other "hidden" conditions, such as neuropsychiatric, metabolic, renal, or other comorbidities that might be simultaneously associated both with altered sleep patterns [34], [35], [240] and an increased risk of cardiovascular events [241]–[244].

Table 7.3 and Supplementary Table C.12 further present the generalization performance of the final RSF with and without the AHI predictor, respectively, on remaining test cohorts of subjects without prior cardiovascular events. The results are stratified by medication use (M = 1) and indication of out-of-domain subjects (†), unseen by RSF during their training. In all scenarios, the models with and without AHI achieved stable and comparable performance across cohorts, with C-index values ranging 66.6–79.3% and 66.6–78.3%, IBS values 4.1–12% and 4.1–11.9%, 1-year tdAUROC 64.5–85% and 64.1–85.2%, 5-year tdAUROC 69.2–84.5% and 69.2–83.4%, and 10-year tdAUROC 69.4–100% and 69–100%, respectively. Strong discriminative performance was further supported by significant differences in predicted mortality between event-free individuals and those who developed cardiovascular events, as determined by a t-test and a log-rank test comparing high- and low-risk groups, in both RSF models and all scenarios. This suggests that RSF performance in subjects without prior events is robust to domain shifts, as medication-taking subjects are typically older, and to potential alterations in sleep induced by medications.

Supplementary Tables C.13 and C.14 present results of the final RSF with and without the AHI predictor, respectively, for baseline (SHHS1) and follow-up (SHHS2) subjects who had experienced at least one cardiovascular event before the corresponding PSG assessment. These individuals were generally older and exhibited higher incidence rates (in their case, recurrence) of cardiovascular events. For comparison, the mean (standard deviation) age in the primary SHHS1(E = 0, M = 0) cohort used for RSF training was 59.4 (11.2) years with a 12.6% incidence (=326/2579) of cardiovascular events, whereas SHHS1(E = 1, M = 0) subjects' age was 68.6 (11.9) years with an incidence of 53.6%, and SHHS1(E = 1, M = 1) had subjects' age of 70.7 (9.7) years with an incidence of 55.9%. The SHHS2 cohorts were, on average, even older. Despite these differences, predicted cardiovascular risk remained informative in these populations (with C-index and tdAUROC exceeding 50% and IBS below

50% in most cases). However, overall discriminatory ability was notably reduced. This decline in performance is not unexpected, as the RSF models were trained on a considerably younger, event-free population and are unlikely to generalize well to post-event subjects whose sleep patterns are likely to substantially differ [235]–[238].

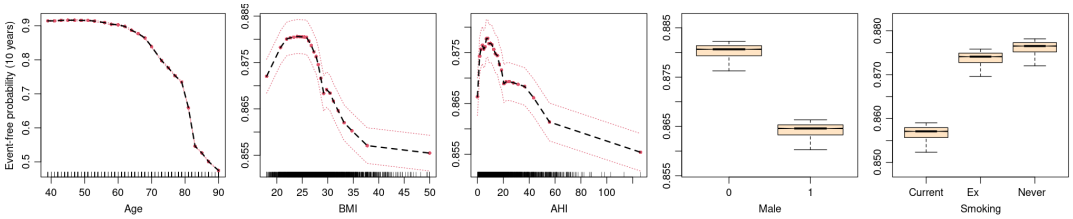
Finally, Figures from Supplementary Sections C.4 andC.5 depict the distribution of cardiovascular cases, survivors, and censoring, as well as the performance metrics (tdAUROC, IBS) for individual study subgroups concerning study wave (SHHS1, SHHS2), medications, prior cardiovascular events, and training vs. unseen subjects, for RFS with and without AHI, respectively.

Cardiovascular risk-profiles via partial effects of RSF

To examine associations between risk factors (demographics, BMI, smoking), sleep patterns (macrostructure, dynamics), and long-term cardiovascular outcomes, we analyzed partial effects from the RSF model, including AHI as a predictor. This adjustment accounts for SDB (AHI) and reduces potential bias. For comparison, we also assessed partial effects from the RSF model without AHI, which showed similar predictive performance.

Figure 7.4 shows the partial effects of demographic factors (age, gender), lifestyle variables (BMI, smoking status), and SDB severity (AHI) on the 10-year cardiovascular event-free probability, interpreted as the complement of risk. The event-free probability remains stable until ∼age 55, then declines, with a steep drop beyond age 80. BMI and AHI exhibit similar risk profiles due to their strong correlation. The highest event-free probability is observed for BMI values between 22–27; risk increases by ∼1% at lower BMI (possibly linked to smoking-related leanness) and up to 3% for BMI>35. A minimal cardiovascular risk is associated with an AHI of ≤15, supporting the clinical threshold that distinguishes mild from moderate SDB. Interestingly, near-zero AHI values are associated with a slight increase in risk, possibly due to poor sleep efficiency (low TST, high WASO), which can reduce AHI yet elevate cardiovascular risk. Partial effects also suggest a 1.5% higher risk in males and a ∼2% increase for current smokers versus never-smokers. Ex-smokers show a modest 0.5% higher risk than never-smokers but 1.5% lower risk than current smokers. These associations align with established evidence linking age, obesity, SDB, and smoking to elevated cardiovascular risk [4], [212]–[214].

Figure 7.4: Partial effects and their 95% CIs for 10-year cardiovascular event-free probability for the age in years, Body Mass Index (BMI), Apnea-Hypopnea Index (AHI), gender (0 = female, 1 = male), and smoking status, for RSF with AHI predictor.

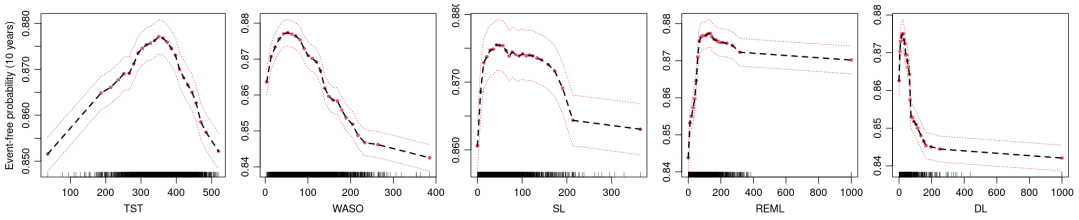


Notes: Data points are shown as ticks on the x-axis.

Figure 7.5 shows partial effects of sleep macro-architecture markers on 10-year event-free probability. Unlike Figure 7.2, where these predictors (except sleep latency, SL) exhibited monotonic, sigmoidal risk profiles for SDB, their associations with cardiovascular risk are predominantly U-shaped. This pattern suggests a "healthy optimum" range with maximal event-free probability, beyond which risk increases bidirectionally. Optimal TST appears around 360 minutes, with both short and long durations, reflecting insufficient sleep and hypersomnia, respectively, linked to up to a 3% increase in cardiovascular risk at the extremes. Short TST may arise from socially or psychiatrically induced sleep deprivation, renal dysfunction, or medical conditions. At the same time, long TST may result from central disorders of hypersomnolence, chronic medical conditions like metabolic dysfunction, systemic inflammation, neurodegeneration, or frailty [41]–[43], all of which contribute to cardiovascular vulnerability. The lowest risk occurs at WASO of 50–80 minutes. Both very

low and very high WASO are associated with elevated risk: low WASO (∼1.5% risk increase) may indicate sleep deprivation, frailty, advanced age, or neurodegenerative changes, while extremely high WASO (>200 minutes) suggests severe sleep fragmentation that may originate in insomnia, depression, chronic pain, or SDB. Sleep latency (SL) below 20 minutes or above 120 minutes corresponds to a ∼1.5% higher risk. Prolonged SL may reflect hyperarousal, insomnia, or anxiety, whereas very short SL could signal excessive sleepiness due to sleep deprivation (e.g., due to long-term SDB); however, SL can also be affected by discomfort or unfamiliarity with PSG recording. Very short REM latency (REML) is associated with a ∼3.5% risk increase, potentially reflecting sleep deprivation, narcolepsy, depression, or metabolic dysregulation. Prolonged REML (>200 minutes) shows a modest risk increase, possibly driven by SDB [39], [222]. Similarly, both very short and prolonged deep-sleep latency (DL) are associated with higher risk (∼1.5% and ∼3.5%, respectively), likely reflecting chronic sleep deprivation, impaired sleep homeostasis, or fragmentation from SDB and depression [38], [219].

Figure 7.5: Partial effects and their 95% CIs for 10-year cardiovascular event-free probability for the minutes of Total Sleep Time (TST), Wake After Sleep Onset (WASO), Sleep Latency (SL), REM Latency (REM), and Deep-sleep Latency (DL), for RSF with AHI predictor.



Notes: Data points are shown as ticks on the x-axis.

Figure 7.6 shows the partial effects of individual sleep-stage transition proportions $p_{i,j}$ (cf. *Methods*), computed relative to the total number of sleep-stage epochs after sleep onset. On average, a post-onset PSG spans ∼910 epochs, so a 1% (= 0.01) change in transition proportion corresponds to roughly 9 transitions per PSG. We focus on clinically relevant transitions with partial effects linked to ≥2% changes in predicted event-free probability, supported by non-overlapping 95% CIs.

- *W-transitions*: The trend in $p_{W,W}$ mirrors WASO, with continuous wakefulness beyond 50% linked to a 4% increase in cardiovascular risk. The optimal range for $p_{W,N1}$ (a marker of fragmentation) is 0.01–0.03, with risk rising by 2% above 0.1. Direct transitions from W to N2, bypassing normal N1 sleep initiation, increase risk by 2% beyond 0.04. Notably, $p_{W,N3} > 0.004$ (∼3 transitions) may signal severe homeostatic sleep pressure and associate with up to a 6% risk increase.

- *N1-transitions*: Frequent N1→W transitions ($p_{N1,W} > 0.02$) reflecting heightened sleep instability, rise risk by >2%. Prolonged periods of light sleep ($p_{N1,N1} > 0.15$, ∼1 hour) show a linear association with ∼2% higher risk. Notably, even a single N1→N3 transition ($p_{N1,N3} \sim 0.001$) links to >6% risk, suggesting abnormal homeostatic responses. Further, $p_{N1,REM} > 0.01$, reflecting atypical REM initiation, shows exponential risk increases, reaching 2% risk increase.

- *N2-transitions*: Cardiovascular risk rises sharply (>7%) when $p_{N2,N1} > 0.004$ (∼3 transitions), indicating instability in intermediate NREM state. The lowest risk occurs with uninterrupted N2 sleep comprising 30–50% of post-onset time; deviations increase risk by >3%, particularly when $p_{N2,N2} > 0.6$, possibly reflecting reduced progression into N3 or REM sleep. Both insufficient ($p_{N2,N3} < 0.005$) and excessive ($p_{N2,N3} > 0.06$) deep sleep transitions are associated with an elevated risk, reflecting disturbed sleep pressure.

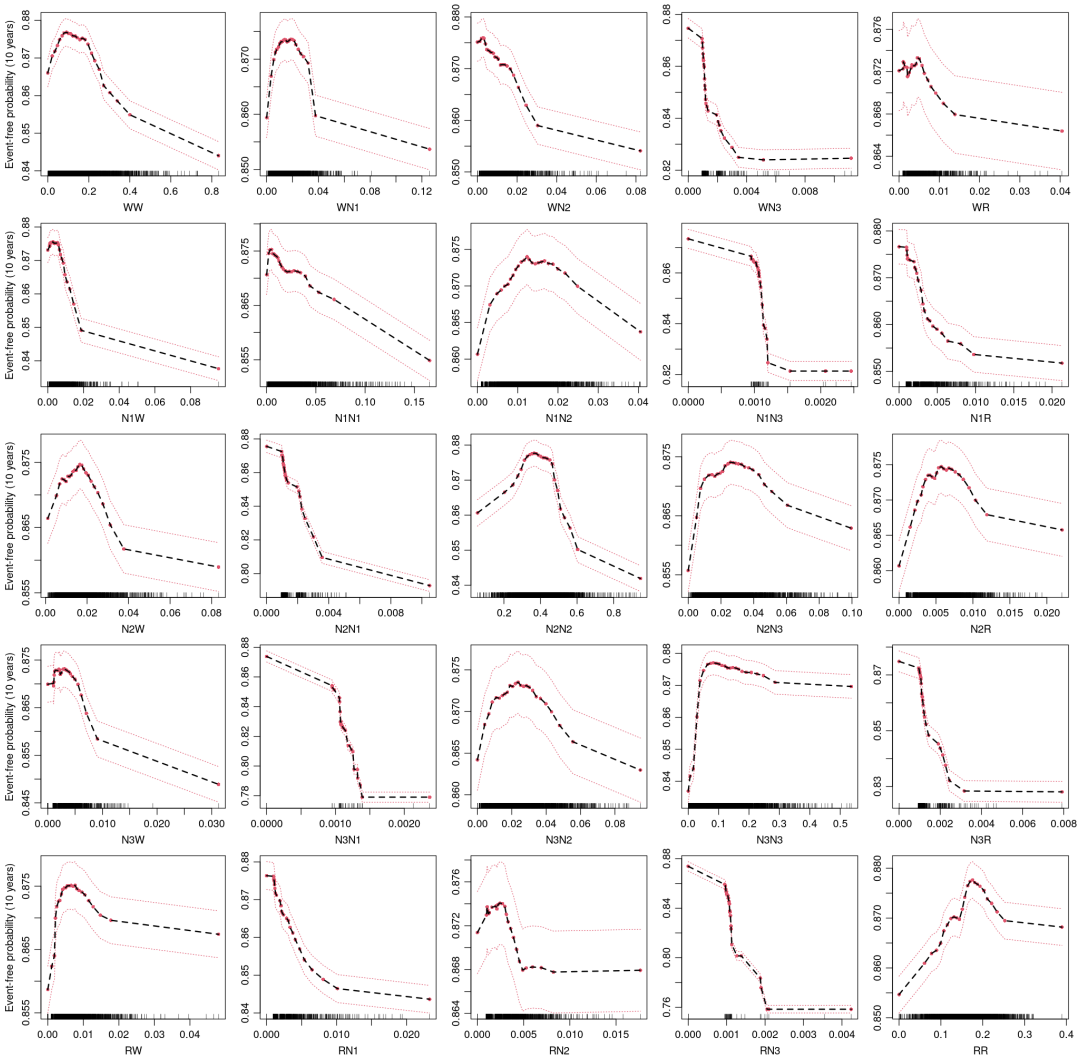
- *N3-transitions*: Deep-sleep disruptions strongly predict cardiovascular risk. N3-awakenings, $p_{N3,W} > 0.1$ increase risk by ∼2%. Atypical $p_{N3,N1} > 0.001$ transitions

link to up to 10% higher risk per single occurrence. Minimal continuous N3 sleep ($p_{N3,N3} \approx 0$) increases risk by 4%, while rare $p_{N3,REM} > 0.001$ associate with ∼4% risk increase.

- *REM-transitions*: REM awakenings, $p_{REM,W}$, outside the [0.05, 0.15] range increase risk by up to 2%. The $p_{REM,N1} > 0.01$ transitions exhibit exponential risk growth (up to 3%), indicating heightened cortical arousals. Atypical $p_{REM,N3}$ are linked to 8–12% higher risk for one or two nightly events, respectively. Optimal continuous REM sleep ($p_{REM,REM}$) lies between [0.15, 0.25] corresponding to ∼1.5 hours of uninterupted REM over the night, while deviations are associated with >2% risk. The REM sleep disruptions likely reflect autonomic imbalance, mood-related dysregulation, or compensatory mechanisms.

By comparing the partial effects from the RSF and RF models, we observe that while SDB risk is primarily associated with monotonic trends (e.g., lower TST, REM, N3, and higher WASO), cardiovascular risk often exhibits non-linear patterns. This likely reflects the influence of diverse clinical conditions beyond SDB that alter sleep parameters at both extremes. Finally, Supplementary Figures C.25-C.27 display partial effects from the RSF model without the AHI predictor, showing trends consistent with those described above.

Figure 7.6: Partial effects and their 95% CIs for 10-year cardiovascular event-free probability for the relative frequencies of transitions between sleep-stage (W, N1, N2, N3, REM), for RSF with AHI predictor.



Notes: Each subplot's x-axis label indicates the direction of the transition (e.g., WN1 corresponds to transitions from W to N1). Data points are shown as ticks on the x-axis.

Correlation of predicted cardiovascular risk with sleep disorders and non-sleep comorbidities

The generalization tests above suggest that our RSF model accurately predicts cardiovascular risk, even under domain shifts involving older individuals and those on confounding medications. Leveraging the clinically rich BSWR data set, which spans a full spectrum of sleep disorders and selected non-sleep comorbidities, we next investigated how these conditions correlate with predicted cardiovascular risk. Each sleep disorder exhibits a distinct pattern of sleep disruption, which may differentially influence cardiovascular risk. Our goal was to determine whether specific sleep disorders and comorbidities are systematically associated with changes in predicted risk. Identifying such associations could provide insight into the cardiovascular relevance of individual sleep pathologies and help highlight high-risk patient subgroups.

Supplementary Table C.2 summarizes the prevalence of major sleep-disorder classes and specific clinical conditions across the BSWR cohort and their associations with moderate-to-severe SDB (AHI $>$ 15). Supplementary Table C.3 provides further details on the demographic and clinical profiles (gender, age, BMI, and AHI) of each condition, comparing them to those of strictly healthy individuals. All seven major classes of sleep disorders significantly differed from healthy controls, showing higher mean age, BMI, and AHI, suggesting that individuals with sleep disorders tend to present with a riskier cardiovascular profile. While the healthy group comprised mostly women (58%), all sleep disorder categories—except insomnia and hypersomnia—had a significantly higher proportion of men ($\geq$61.9%).

Across all major sleep disorder classes, we observed higher average predicted cardiovascular risk. However, this increase cannot be directly attributed to the disorders themselves, as individuals with sleep disorders also differed significantly from controls in demographics, BMI, and AHI. To address this, we quantified the adjusted risk using logistic regression models with log-odds-transformed predicted risk as the outcome, adjusting for age, gender, BMI, AHI, and a binary indicator for the specific condition. Each model was estimated in a case-control design contrasting healthy individuals with those affected by a given condition. The adjusted effect, reported in the final column of Supplementary Table C.3, represents the systematic percentual increase in predicted cardiovascular risk relative to healthy controls, after accounting for differences in demographics, BMI, and AHI.

Even the adjusted models revealed significant risk increases across all major sleep disorder classes: SDB was associated with a 17.8%, 95% CI: (9.5, 26.8), increase in risk, insomnia 12.4% (2.9, 22.7), hypersomnia 11.0% (3.7, 18.9), movement-related disorders 20.3% (8.6, 33.2), parasomnias 24.5% (13.3, 36.9), circadian rhythm disorders 27.5% (10.3, 47.3), and isolated symptoms or normal variants 10.9% (2.9, 19.5). Specific conditions within these categories exhibited variability, as evidenced by a 39.2% (23.5, 56.9) increase for central sleep apnea, a 44.4% (27.0, 64.1) increase for NREM parasomnias, and non-significant effects for narcolepsy type 2, short-term insomnia, and idiopathic hypersomnia. Among comorbidities, the largest risk increase was observed for neurodegenerative diseases, 45.8% (21.4, 75.1); diabetes, 39.4% (13.9, 70.7); and current cardiac disease or prior event, 25.1% (8.7, 44.0).

These observed associations between sleep disorders, comorbidities, and elevated cardiovascular risk underscore the interrelationship between sleep and cardiovascular health. They also suggest that distinct patterns of sleep disruption, attributable to specific conditions, elevate the cardiovascular risk at different levels. However, further studies in cohorts with long-term monitored cardiovascular outcomes are needed to validate these observations, as such data are not currently available in BSWR.

7.4 Discussion

A wide range of clinical conditions, including both sleep disorders and non-sleep comorbidities, can disrupt sleep macrostructure by altering total sleep duration, increasing fragmentation, and modifying the distribution of sleep stages [8], [13], [36], [165], [245], [246]. In particular, SDB, affecting up to 23.4% and 49.7% middle to older-aged women and men [231], respectively, induces characteristic macrostructural changes, including reductions in REM and N3 sleep, increased fragmentation, and altered stage composition [25], [26], [218], [219].

SDB has also been linked to elevated cardiovascular morbidity and mortality, including hypertension, stroke, and sudden cardiac death [2], [4], [5], [7], [212]–[216]. Statistically, if a condition such as SDB alters both sleep macrostructure and the cardiovascular health, macrostructural sleep patterns may carry predictive information about cardiovascular risk, at least in an associative sense. Supporting this, reduced total sleep duration, as well as lower proportions of N3 and REM sleep, have been associated with increased cardiovascular risk [37]–[43]. Emerging evidence suggests that sleep-stage dynamics—the temporal patterns of transitions between stages—may offer deeper insights into physiological regulation and disease-specific signatures [26], [100], [105]–[107], [109], [115], [171]. While prior studies focused on static macrostructural features, the prognostic relevance of dynamic sleep patterns for cardiovascular outcomes remains largely unexplored. Since alterations in sleep dynamics, like those in macrostructure, often arise as downstream effects of conditions such as SDB, diabetes, chronic pain, and neurodegenerative disorders, these patterns may encode signals relevant to cardiovascular risk.

Our study leveraged data from the SHHS, a prospective cohort originally designed to investigate the relationship between SDB and cardiovascular risk [182]. We used these data to examine the dependency chain linking SDB (a major cardiovascular risk factor), sleep characteristics (macrostructure, dynamics), common risk factors (demographics, BMI, smoking status), and long-term cardiovascular outcomes. Specifically, we assessed (i) whether SDB can be predicted from sleep parameters and common risk factors, and (ii) whether long-term cardiovascular risk can be predicted using the same features, both with and without explicit knowledge of SDB severity. To quantify these relationships, we applied forest-based methods [224], [227]—Random Forest (RF) for SDB identification and Random Survival Forest (RSF) for cardiovascular risk prediction. These models are well-suited for capturing complex, non-linear relationships, robust to overfitting and multicollinearity, and support interpretability through partial dependence analyses. Notably, we analyzed R(S)F partial effects to determine whether individual predictors exhibited predominantly linear associations, as often assumed in prior studies using restrictive methods such as ANOVA [37] or regression-based models [38]–[43], or displayed non-linear patterns suggesting ranges of clinical optima with minimal risk. All models were trained on a carefully stratified SHHS baseline cohort, which was free from prior cardiovascular events and medication use, thereby minimizing confounding and enhancing the generalizability of our findings to a broader population.

The RF demonstrated that SDB can be reliably detected from the considered predictors. Cross-validation in the primary study cohort yielded an AUROC of 76.1% for identifying moderate-to-severe SDB (AHI >15), with strong generalization across REM- and NREM-dominant phenotypes (74.1-74.9%) and mixed SDB (79.4%). SDB detection remained robust even in unseen subgroups with prior events or medications from SHHS1 (73.6–74.1%), SHHS2 follow-up cohorts (69.5–80.6%), and in a fully out-of-domain clinical BSWR (76.0%). These findings demonstrate that SDB can be reliably inferred from sleep parameters and common risk factors only, even in medication or prior-event-confounded subgroups, and without direct access to respiratory signals typically required for clinical diagnosis. Partial dependence analyses revealed predominantly monotonic trends, with SDB risk increasing sharply above age 50, BMI >25, in males and ex-smokers, consistent with existing evidence [5], [213], [214], [218], [230]–[233]. Macrostructural sleep markers of SDB included TST <300 minutes, WASO >100 minutes, and prolonged REM and N3 latencies, confirming that apneic events cause fragmented and inefficient sleep, with delayed progression into restorative states [106], [218], [219], known to be important for brain recreation. Novel insights emerged from sleep-stage transition proportions ($p_{i=\text{from},j=\text{to}}$) proposed in our previous work [115], where $p_{i,j} = 0.01$ corresponds to roughly nine transitions per night while $p_{i,i} = 0.1$ indicates about 45 minutes of uninterrupted time in stage i. Several transitions proved to be highly sensitive markers, associated with >5% increases in SDB risk, including $p_{W,N2} > 0.02$, $p_{W,REM} > 0.01$, $p_{N1,W} > 0.02$, $p_{N2,W} > 0.02$, and reduced REM continuity ($p_{REM,REM} < 0.1$). While prior studies have quantified the effects of how the SDB alters sleep macrostructure and dynamics [26], [105], [106], [109], [115], our findings suggest that these patterns alone enable effective screening of SDB, with partial effects providing mechanistic insight into these associations.

The RSF models further quantified the extent to which cardiovascular risk can be stratified from the same set of predictors. Two versions were trained: one that included SDB

severity (AHI) and one that excluded it. Strikingly, inclusion of AHI did not improve performance on any discrimination or calibration metric assessing predictive capability to capture cardiovascular risk. For example, the cross-validation yielded C-indices of 73.0% and 73.3%, 10-year tdAUROCs of 75.1% and 75.3%, and IBS values of 6.7% for models with and without AHI, respectively, with significant log-rank tests having p-values of the same order. These findings suggest that in a flexible-enough model (RSF), demographic factors, BMI, smoking status, and sleep parameters sufficiently capture pathological signatures of cardiovascular risk, to the point that adding AHI offers no additional predictive benefit. This likely reflects the ability of these predictors to encode not only SDB-related patterns (as shown by the SDB-identification experiment) but also other pathological, possibly undiagnosed processes—such as diabetes [241], renal dysfunction [242], cancer [247], pain syndromes [244], or neurodegeneration [243]—that may jointly influence sleep and cardiovascular outcomes.

Supporting this, RSF partial effects of individual predictors revealed non-linear, often U-shaped risk profiles (in contrast to the monotonic effects in RF for SDB detection), suggesting clinical optima and thresholds of increased cardiovascular risk. Minimal risk was observed for age under 55 years, BMI $\in [20, 25]$, AHI <15, and never-smokers. Macrostructural markers of minimal risk included TST $\in [300, 400]$ minutes and WASO $\in [40, 100]$ minutes, while deviations from these ranges—along with excessively short or long sleep-onset, REM, and N3 latencies—were associated with higher risk. Sleep-stage continuities in N2, N3, and REM stages exhibited protective ranges at $p_{N2,N2} \in [0.3, 0.5]$, $p_{N3,N3} \in [0.1, 0.3]$, and $p_{REM,REM} \in [0.15, 0.25]$, corresponding to about [135, 225], [45,135], and [67.5, 112.5] minutes, respectively. These U-shaped risk profiles confirm prior associations between reduced TST, N3, and REM sleep durations with cardiovascular morbidity and mortality [38]–[43], while extending them by showing that risk also increases above optimal values—a nuance not captured in earlier studies constrained to linear models. In addition, rare or highly atypical transitions, seldom observed in healthy sleep, were strongly associated with sharply monotonically increased risk. For instance, >3% risk increase was linked to $p_{W,N3} > 0.002$, $p_{N1,W} > 0.03$, $p_{N1,N3} > 0.001$, $p_{N2,N1} > 0.003$, $p_{N2,N2} > 0.6$, $p_{N3,N1} > 0.001$, $p_{N3,N3} < 0.05$, $p_{N3,REM} > 0.002$, $p_{REM,N1} > 0.0075$, $p_{REM,N3} > 0.001$, and absence of continuous REM sleep ($p_{REM,REM} \approx 0$). Notably, even a single occurrence of such atypical transitions (e.g., $p_{N1,N3}$, $p_{N3,N1}$, $p_{REM,N3}$) during a night may serve as a sensitive marker of cardiovascular risk, whether driven by SDB or other underlying conditions. Our findings extend the existing knowledge that sleep dynamics are not only useful for describing present clinical conditions, but also provide signals correlating with future health events.

Movel validation confirmed strong generalization of RSF predictions across SHHS subgroups with medication use (C-index >66.6%, IBS $\leq$12%, tdAUROC >69%, and significant log-rank test in all baseline or follow-up subgroups). However, performance was reduced in subjects with prior cardiovascular events, likely due to altered sleep-wake patterns caused by events (cf. [235]–[238]) and also much older age. In the BSWR data set, predicted cardiovascular risk was positively associated with all seven major sleep disorder classes (SDB, insomnia, hypersomnia, parasomnias, movement disorders, circadian-rhythm disorders, and isolated symptoms), with estimated adjusted increases in cardiovascular risk ranging from 10.9% to 27.5% compared to healthy controls. Among non-sleep comorbidities, neurodegenerative diseases, diabetes, and existing cardiac disease were associated with the highest increases (>25%). These findings collectively support the strong interplay between sleep and cardiovascular health.

7.5 Conclusion

Our study demonstrates that sleep macrostructure and stage dynamics jointly encode sensitive markers of both current SDB and long-term cardiovascular risk. Leveraging carefully curated data from the large prospective SHHS cohort, we show that SDB can be reliably identified from sleep patterns and demographics alone, without the need for direct respiratory measurements. While we confirm established associations between short duration of TST, REM, and N3 with cardiovascular risk, the use of a flexible RSF modelling approach uncovered non-linear U-shaped relationships, revealing that excessive amounts of specific sleep

features (including TST, REM) are also linked to increased risk—patterns overlooked by traditional linearly restricted methods. Notably, partial effects—providing insights into associations between risk and individual predictors—were largely monotonic for SDB, whereas cardiovascular risk exhibited predominantly U-shaped profiles, suggesting distinct physiological mechanisms and thresholds that could serve as novel markers in clinical decision-making. This suggests that cardiovascular vulnerability involves broader processes beyond SDB, reflected as downstream effects encoded in disrupted sleep. Hence, sleep architecture and dynamics act as a mirror of health, capturing signatures of current physiological states and predicting future disease risk. Together, they position sleep-stage patterns as promising, non-invasive biomarkers for diagnosing current conditions and stratifying long-term cardiovascular risk. With the rise of wearable technologies and automated sleep scoring, combined with additional biosignals such as respiratory patterns, oxygen saturation, and heart rate, these insights highlight the potential for large-scale, unobtrusive, and long-term monitoring, as well as future screening tools for cardiovascular health.

7.6 Limitations

This study has several limitations. Despite using partial dependence analysis in the R(S)F framework, the quantified effects should be interpreted cautiously, as the modelling approach captures numerical associations and hence, partial effects should not be viewed as causal. In addition, as altered sleep patterns likely reflect downstream effects of different underlying conditions, the treatment interventions should target the root causes (e.g., SDB, diabetes management) rather than modifying sleep parameters in isolation. Next, our models were trained on participants free from prior cardiovascular events and medications, which, although improving generalizability, limits applicability in these subgroups. While forest-based methods can internally handle interactions between predictors, incorporating explicit age–gender interactions may be valuable, particularly given the protective effect of pre-menopause on cardiovascular outcomes. Additionally, we modelled a pooled composite cardiovascular endpoint, which may obscure specific risk patterns for individual outcomes. Future work could leverage competing risk models to disentangle and address each cardiovascular event separately. External validation of cardiovascular risk predictions in the BSWR data set was limited to adjusted associations with clinical conditions due to the lack of standardized time-to-event data. Therefore, in our future work, we plan to integrate causal knowledge on clinical relations, expand modelling to include key pharmacological classes and menopausal status, and harmonize data across multiple cohorts tracking long-term outcomes to enable robust cross-cohort validation.

Chapter 8

Discussion

This final chapter synthesizes the six manuscripts presented in the preceding Chapters 2-7 and summarizes their main contributions, including both technical advances and clinical implications. In addition to outlining these contributions, it also discusses the limitations of the individual studies and of the thesis as a whole, and considers how these may be addressed in future work.

8.1 Summary of research findings

In line with the Introduction Chapter 1, the main contributions of this dissertation are grouped into two thematic branches. The first branch, *Integration of Automated Sleep Scoring into Clinical Practice*, encompasses Chapters 2–4 and addresses ethical and legal requirements for deploying AI in healthcare, with a focus on human oversight (Chapter 2) and algorithmic fairness (Chapters 3–4) in the context of sleep scoring. The second branch, *Digital Biomarkers from Sleep-Stage Dynamics*, spans Chapters 5–7 and applies explainable machine learning to uncover novel insights into sleep-disordered breathing, chronic fatigue and pain syndromes, and long-term cardiovascular outcomes.

8.1.1 Integration of Automated Sleep Scoring (ASS) into Clinical Practice

Chapter 2: *Bridging AI and Clinical Practice: Integrating Automated Sleep Scoring with Uncertainty-Guided Physician Review*

This study addressed the challenge of aligning AI predictions with physician responsibility in sleep scoring [95]. Motivated by the fact that inter-scorer agreement between human experts typically ranges from 75–85% [14], [15], [48], [65]–[69], ASS algorithms trained on large, heterogeneous datasets containing scoring patterns of multiple experts achieve comparable levels in performance metrics [80], [141]. As a consequence, approximately 15–25% of epochs remain discordant between algorithmic and human scoring, even for state-of-the-art systems. When deploying ASS in clinical practice, efficient mechanisms for human oversight are therefore essential, as unguided review can take nearly as much time as manual scoring from scratch, limiting its clinical utility. Whereas most prior work has focused on optimizing performance metrics of ASS algorithms, relatively few studies have investigated their effective deployment in human-in-the-loop pipelines. Uncertainty estimation has been proposed as a potential solution, but existing approaches typically rely on functions of the predicted probabilities (i.e., softmax scores) [64], [93], [94], [122]–[124], and have rarely been evaluated for clinical usability across diverse sleep disorders.

In this work, we systematically compared several softmax-based uncertainty metrics and further introduced a novel LSTM-based auxiliary confidence network. Unlike softmax-only approaches, this network integrates both the softmax outputs and representations from intermediate layers of the deep-learning classifier U-Sleep, thereby leveraging sequential features of PSG data. In identifying predictions likely to disagree with expert scorers, the confidence network outperformed all softmax-based baselines across both the in-domain and two out-of-domain test sets scored by individual senior physicians (AUROC $\geq$82.5% in all cases). The ability to accurately flag potentially misclassified epochs is a prerequisite for establishing efficient human oversight. Through simulated querying of predictions below varying confidence thresholds, we demonstrated that revising fewer than 29% of the least

confident stages was sufficient to reach near-perfect agreement between ASS system U-Sleep and physicians ($\kappa \geq 0.90$). Moreover, the in-depth evaluations revealed that predicted confidence scores were significantly lower for algorithm–human disagreements, and that subject-level mean confidence scores positively correlated with classification performance metrics, supporting their interpretability at both the epoch and subject levels, independent of sleep-disorder status.

Together, these results demonstrate that incorporating uncertainty estimation provides a practical and effective mechanism for human oversight, substantially reducing the burden of review, and hence also the overall cost of PSG assessment, while maintaining clinical reliability. Unlike earlier studies limited to restricted cohorts, our work provides the first comprehensive evaluation of uncertainty-based approaches in ASS across a full spectrum of sleep disorders, within both in-domain and out-of-domain tests, demonstrating both the superiority of the proposed confidence network and its robustness across diverse patient populations.

Chapter 3: *Framework for Algorithmic Bias Quantification and its Application to Automated Sleep Scoring*

In this study, we developed a general framework for quantifying algorithmic bias that is applicable to any predictive model in a regression setting [96]. Existing validation methods typically rely on correlation analyses, which measure linear association between predicted and reference values, or Bland–Altman (BA) plots, which assess the magnitude of prediction errors relative to reference values [134]. While informative, both approaches rely on restrictive assumptions such as linearity and homoscedasticity of errors, and they overlook the potential influence of external factors such as demographic or clinical characteristics [136].

To overcome these limitations, our study proposed to model the systematic error (*bias*), defined as prediction–reference differences, possibly conditional on external factors (*sensitive attributes*), using Generalized Additive Models for Location, Scale and Shape (GAMLSS) [138]. Within this framework, the systematic error is captured using an extended normal distribution with separate predictors for expectation (location) and variability (scale). This setup allows flexible modelling of both the mean error and its dispersion, and enables nonlinear effects of external factors (e.g., age) to be captured through splines or more complex predictor bases such as neural networks. Once estimated, the bias model supports hypothesis testing of factor-specific effects (i.e., factor-driven biases), estimation of arbitrary conditional quantiles (e.g., 5% worst- or best-case scenarios), and calculation of coverage within a clinically defined Region of Practical Equivalence (ROPE).

As a use case, we applied the framework to the state-of-the-art deep-learning-based ASS algorithm U-Sleep [59], [60], evaluated on 4,075 PSGs from the Bern Sleep-Wake Registry. Whereas most ASS studies report only epoch- or subject-level classification metrics, few assess the validity of sleep-scoring–derived clinical markers, despite their central role in diagnostics and decision-making. We therefore focused on wake percentage (W%), which directly reflects the ability of the ASS system to distinguish sleep–wake states and underpins the calculation of sleep efficiency and related indices such as TST, WASO, and awakening rate [8]. The bias was model under consideration of the spline effect for the age, and linear effects of gender, AHI, and PLMI, for both location and scale distributional parameters. Most importantly, the analysis revealed systematic, nonlinear age effects on W% errors: U-Sleep consistently underestimated W% in children (median bias of up to -8% in newborns), with both bias magnitude and variability highest at the youngest and oldest age ranges. These findings reflect imbalances in the original U-Sleep training data and illustrate how the framework can uncover clinically meaningful biases as well as technical insights.

Together, this work introduces a practical and flexible methodology for detecting and quantifying bias in predictive algorithms. While illustrated on ASS, the framework is broadly applicable to other systems where predictive numerical outputs underpin clinical, scientific, or industrial decision-making (e.g., apnea detectors predicting AHI values, wearables assessing blood saturation, pricing models). By explicitly modelling bias distributions conditional on arbitrary external factors, it enables a transparent assessment of fairness and reliability, in line with regulatory requirements such as the EU AI Act, MDR, and MedDO. Crucially, as illustrated in our use case, it may also highlight gaps or imbalances in training

data, providing actionable guidance for data curation or stratified model retraining. The ability to quantify potentially nonlinear error distributions dependent on external factors makes it suitable for clinical certification assessments of predictive software tools.

Chapter 4: *Beyond Accuracy: Extending Bias Quantification to Performance Metrics and Clinical Markers*

In this follow-up study [97], we extended the bias quantification framework from Chapter 3 to also cover algorithmic classification performance metrics, which are typically bounded between 0–1 (0–100%). For this purpose, we proposed to employ the zero-and-one-inflated Beta distribution within the GAMLSS framework [138], [148]. This choice enables flexible modelling of both the central tendency and variability of performance metrics, possibly depending on external factors, while accounting for extreme cases of perfect (1) or failed (0) classification. Such an approach allows a more comprehensive assessment of model validity and capability across, e.g., demographic and clinical, subgroups, directly addressing concerns about fairness and equity of predictive models in healthcare.

As a use case, we compared two widely used ASS systems: the state-of-the-art deep-learning algorithm U-Sleep [59], [60] and the ML-based YASA [58]. The evaluation included a wide set of hypnogram-derived clinical markers (e.g., TST, WASO, REML) where biases were assessed using extended normal distribution (cf., Chapter 3) together with two standard performance metrics: macro-F1 and accuracy, using an inflated Beta distribution. Results revealed systematic differences in biases between the algorithms, with nonlinear age effects and linear worsening effects of AHI and PLMI. These findings suggest that both demographic and clinical factors substantially affect model performance and should be considered in fairness assessments, or when using these ASS tools in clinical decision-making.

Importantly, the observed age-related biases in both clinical markers and performance metrics may reflect the absence of age as an explicit input variable of ASS systems, related to the omitted variable bias phenomenon. As both sleep architecture [118], [156] and raw PSG biosignals [119], [157] evolve with age, algorithms lacking age information may struggle to learn these interactions, and equal performance across age groups cannot be guaranteed. Despite that, vast majority of ASS tools ignore age as its imput [58], [59], [73].

Furthermore, our study assessed whether biases in hypnogram-derived markers lead to reduced diagnostic value. Using obstructive sleep apnea (OSA) as an example, we demonstrated that when predictive errors are consistent (i.e., systematic biases), simple machine learning classifiers such as LASSO logistic regression or Random Forest can adapt to them and achieve comparable performance regardless of whether they were trained on reference markers derived from physician scoring or on biased predictions from ASS. This illustrates that, on the one hand, predictions must be treated with caution and carefully validated, yet on the other hand, they may still provide valuable information despite their inherent biases.

Together, the extension of the bias-quantification framework to performance metrics demonstrated that the fairness and reliability of ASS (or other predictive) systems cannot be fully assessed through mean-level performance summaries alone. By modelling full performance distributions, the framework provides subgroup-specific insights into systematic strengths and weaknesses of algorithms, offering practical guidance for data curation, model retraining, and regulatory evaluation.

In summary, automated sleep scoring (ASS) has reached a level where it provides clinically meaningful insights but remains constrained by the inter-scorer variability of human experts, which limits achievable performance. Physicians must therefore remain the final decision-makers, supported by mechanisms for effective human oversight (Chapter 2). To ensure fair, transparent, and clinically reliable use of ASS, it is also crucial to understand algorithmic behaviour and potential biases in both clinical markers and performance metrics (Chapters 3-4). Since such variability is intrinsic to human scoring, the integration of ASS into clinical workflows is best supported by transparent disclosure of the underlying scoring mechanisms, their uncertainties, and potential biases. Along these lines, recent benchmarking initiatives, most notably SLEEPYLAND [73], have created a platform allowing transparent and standardized comparisons of state-of-the-art ASS algorithms on common datasets.

Importantly, this effort has also incorporated the bias-quantification framework developed in Chapters 3-4, highlighting its practical value for promoting transparency and fairness evaluation.

8.1.2 Digital Biomarkers from Sleep-Stage Dynamics

Chapter 5: *Novel Digital Markers of Sleep Dynamics: A Causal Inference Approach Revealing Age and Gender Phenotypes in Obstructive Sleep Apnea*

In this study [115], we developed a framework to quantify novel digital biomarkers of sleep disorders based on sleep-stage dynamics extracted from an observational clinical PSG database. A major methodological challenge in such settings is the presence of confounding: case–control distributions are not randomized, demographic profiles differ substantially, and patients frequently suffer from comorbidities. Except for a few routinely reported indices, such as the number of awakenings or total transition rates, current clinical PSG evaluation does not systematically focus on sleep-stage dynamics, despite their potential to reveal deeper physiological and pathological mechanisms of sleep regulation (cf. [98]–[114]).

Motivated by this gap, our work proposed a simple yet characteristic biomarker of sleep disorders—the raw 5×5 matrix $\mathbf{P}$ of sleep-stage transition proportions—and a methodological approach for its estimation from observational data. Marker $\mathbf{P}$ functionally relates to established hypnogram-derived metrics (e.g., sleep-stage proportions) while indirectly also capturing overall stage durations, thereby linking to previously studied aspects of sleep-stage fragmentation [98]–[107] and continuity [108]–[114].

Our methodological framework combined elements of causal inference to minimize bias in estimating disease effects on marker $\mathbf{P}$. It consisted of two main components: (i) *a propensity score model* estimating disease probability conditional on key confounders, satisfying the positivity assumption, i.e., ensuring sufficient overlap of covariate distributions between cases and controls [177], and (ii) *an outcome model* quantifying the effect of disease on $\mathbf{P}$, corresponding to a causal S-learner [180], while adjusting for interactions with comorbidities and demographics. The causal estimand of interest was the conditional average treatment effect (CATE), representing the difference in expected outcomes (i.e., dynamics captured in $\mathbf{P}$) between cases and controls, personalized to the levels of predictors in the outcome model. Previous studies of sleep disorders have often ignored confounding and typically reported only unadjusted case–control differences in disease effects or markers using simple statistical tests, with very few considering the role of ageing and demographics [9], [17], [166], or sleep comorbidities, despite their high prevalence in clinical populations [190], [191], [240].

We demonstrated this framework in the context of obstructive sleep apnea (OSA), one of the most prevalent sleep disorders, affecting an estimated 17% of the general adult population [149] and up to 23.4% and 49.7% of middle- to older-aged women and men [231], and known to be a major risk factor for cardiovascular morbidity and all-cause mortality [2], [213], [215]. The study dataset, a subset of the Bern Sleep-Wake Registry (BSWR), included 62 healthy controls and 560 OSA cases, with more than 48% of the latter presenting at least one additional sleep comorbidity. Propensity score weighting using logistic regression was applied to balance demographic confounders, and a Dirichlet outcome regression was employed to jointly model all dimensions of $\mathbf{P}$, adjusting for OSA and its interactions with OSA severity (AHI), demographics, and comorbidities. The Dirichlet formulation further enabled aggregation across matrix dimensions, allowing us to derive novel interpretable metrics such as stage-specific fragmentation rates and NREM continuity, complementing existing hypnogram-derived PSG indices. Uniquely, the outcome model fitted the rich BSWR cohort enabled personalized CATE estimation of OSA effects on sleep dynamics, disentangled from the influence of comorbid sleep conditions. To facilitate exploration and broader research community outreach, we additionally developed an interactive web application.

Our main findings indicated that markers of NREM–REM oscillations and stage-specific fragmentation were consistently increased across all OSA severities and demographic groups. Moreover, we identified distinct gender-specific phenotypes, with females exhibiting higher vulnerability to awakenings and REM-related disruptions, which may explain their more frequent reports of insomnia- or depression-like symptoms in OSA [193]–[195].

In summary, this study presented a flexible framework for quantifying novel digital markers of sleep disorders based on sleep-stage dynamics from observational clinical data, and demonstrated its utility in OSA. Beyond establishing new markers, we showed that they are also predictive of disease presence, underscoring their potential for both clinical stratification and mechanistic insights, in line with the overarching goals of this thesis to develop interpretable and clinically meaningful computational tools for sleep medicine.

Chapter 6: *Unveiling Sleep Dysregulation in Chronic Fatigue Syndrome with and without Fibromyalgia Through Bayesian Networks*

In this study [116], we investigated how Chronic Fatigue Syndrome (CFS) and its frequent comorbidity Fibromyalgia (FM) affect sleep regulation and sleep-stage dynamics. Both syndromes are more prevalent in females, share overlapping symptoms such as non-restorative sleep and daytime fatigue, yet differ in clinical presentation: CFS being dominated by exertional fatigue and FM by widespread pain [203], [204]. Their frequent co-occurrence [202] makes clinical differentiation challenging, while standard PSG indices often yield inconsistent results [208]. Sleep-stage dynamics, by characterizing transitions and temporal structure, may therefore provide deeper insights into physiological dysregulation and support clinical distinction.

We analyzed a small but high-quality dataset collected by Kishi et al. [201], comprising PSG recordings from 26 healthy women, 14 with CFS, and 12 with CFS+FM, all aged 25–55. This strictly controlled experimental cohort minimized variability and confounding: groups were demographically matched, participants were free of psychiatric or sleep disorders, instructed to abstain from alcohol, caffeine, and strenuous activity, and recorded during the follicular menstrual phase to reduce hormonal effects. While Kishi et al. identified differences in stage prevalence and first-order transitions using simple statistical tests, these analyses offered only limited insight. To extend this work, we developed a dynamic Bayesian Network (BN) with an expertly informed causal structure to jointly model stage prevalence, bout durations, and transitions, and to quantify the specific impacts of CFS and CFS+FM. The controlled study design minimized confounding and enabled a causal interpretation of the estimated effects.

Our results confirmed that sleep dynamics are best described as a second-order process, with an optimal lag of two previous stages. This finding aligns with reports in general populations [107] and suggests its validity even in our clinical cohort. The final BN achieved robust next-stage predictions with an in-domain accuracy of 70.6% and generalization accuracies of 60.1–69.8% on two independent validation cohorts. It also differentiated healthy, CFS, and CFS+FM subjects with an AUROC of 75.4%. Beyond prediction, we performed simulated interventions by fixing the health status node of BN (healthy, CFS, or CFS+FM) and sampling sleep trajectories under individual conditions. This approach, conceptually similar to do-calculus [210], allowed us to estimate the causal effect of each condition on sleep dynamics. The results revealed prominent alterations, including prolonged wakefulness and N3 durations in both conditions, extended REM bouts in CFS, and reduced N1/N2 durations in CFS+FM. Together, these patterns suggest that CFS is marked by impaired maintenance of restorative REM sleep, contributing to unrefreshing sleep complaints [102], [207], while FM is associated with compensatory increases in deep sleep but instability in cycling, consistent with pain-related sleep disruption [204].

In summary, this study applied Bayesian Networks as an approach that directly and transparently encodes cause–effect structures, providing a complementary tool to the causal inference framework used in Chapter 5 for estimating effects (markers) on sleep-stage dynamics. Using strictly controlled clinical data, the BN confirmed the second-order nature of sleep, achieved strong predictive and diagnostic performance, and quantified disorder-specific alterations through simulated interventions. Importantly, the analysis of first-order transitions confirmed and extended the existing findings [201], while second-order transitions revealed novel alterations not previously described. These estimated effects, together with detailed insights into individual transitions, can be regarded as novel digital markers, as they not only captured disorder-specific dysregulation but also demonstrated diagnostic capability. Methodologically, whereas the OSA study in Chapter 5 captured sleep dynamics via a raw matrix of stage transition proportions, here we assessed them in terms

of stage bouts, their identifiers, and durations, offering a complementary representation of sleep. Collectively, these findings underscore the potential of sleep-stage dynamics as digital biomarkers, supporting clinical differentiation and providing mechanistic insights into complex conditions such as CFS and FM.

Chapter 7: *Sleep-Stage Dynamics Predict Current Sleep-Disordered Breathing and Future Cardiovascular Risk*

In this study [117], we investigated the predictive value of sleep for long-term cardiovascular outcomes, with a particular focus on sleep-stage dynamics. In particular, sleep-disordered breathing (SDB, such as OSA) is a well-established risk factor for cardiovascular morbidity and mortality [2], [7], [212]–[215]. At the same time, SDB is known to disrupt both sleep macrostructure, reducing proportions of restorative N3 and REM sleep and increasing light sleep (N1, N2) and fragmentation [25], [26], [218], [219], and sleep-stage dynamics, i.e., the temporal continuity and organization of transitions between stages [26], [106], [109], [115]. Some prior studies have linked certain macrostructural features, such as reduced slow-wave (N3) and REM sleep, to elevated cardiovascular risk [37]–[40]. However, to date, no work has directly assessed the predictive power of sleep-stage dynamics. Given that dynamics capture finer-grained regulatory signatures of body physiology, they may provide unique prognostic information beyond static macrostructure metrics (cf. [98]–[114]). To test this, we used flexible forest-based models [224], [227], capable of capturing non-linear effects and complex interactions beyond the limitations of regression-based approaches, to jointly assess whether sleep macrostructure, dynamics, and established risk factors (age, BMI, smoking) encode predictive patterns of cardiovascular risk. Our intuition was that, because OSA is both a major cardiovascular risk factor and a condition that profoundly alters sleep architecture, it is reasonable to expect that sleep-stage dynamics would contain predictive signatures of cardiovascular vulnerability.

To assess these relations, we used data from the prospective, community-based Sleep Heart Health Study (SHHS) [182], focusing on 2579 participants without prior cardiovascular events or sleep-altering medications, thereby minimizing confounding and enhancing generalizability. We first applied a Random Forest classifier and demonstrated that moderate-to-severe SDB can be reliably identified from sleep parameters and common risk factors alone (AUROC = 76.1%), with robust generalization across REM-, NREM-, and mixed-dominant phenotypes as well as out-of-domain validation cohorts. Since SDB is a major cardiovascular risk factor, its detectability from sleep patterns implies that the underlying patterns related to SDB detection should also carry prognostic information about cardiovascular outcomes. Consistent with this, our Random Survival Forest predicted long-term cardiovascular risk with strong performance (concordance index = 73.3%, 10-year time-dependent AUROC = 75.3%). Strikingly, adding the apnea–hypopnea index (AHI, i.e., direct measurement of SDB) to the predictor set did not improve model performance, indicating that sleep architecture and dynamics, together with demographics and lifestyle factors, already encode the prognostic information attributable to SDB measurement.

Analysis of partial effects provided mechanistic insights into how individual predictors relate to SDB and cardiovascular risk. For SDB detection, effects were largely monotonic: risk increased steadily with age, BMI, and male sex, and sleep macrostructure showed associations with SDB such as short total sleep time, long wake after sleep onset, and delayed REM or N3 latency. In contrast, cardiovascular risk was characterized by predominantly U-shaped associations, suggesting the presence of "optimal ranges" for many sleep parameters: for example, event-free survival was maximized for total sleep time of 6–7 hours, wake after sleep onset of 50–80 minutes, and intermediate ranges of N2, N3, and REM continuity, whereas both lower and higher values were linked to elevated risk. Importantly, rare and atypical transitions, such as N3→N1, REM→N3, or excessive awakenings from N2, were associated with sharp increases in cardiovascular risk, even if they occurred only once or twice per night. These results extend prior work that identified linear associations between macrostructural sleep features and cardiovascular outcomes [37]–[40], by revealing non-linear patterns and highlighting sleep-stage dynamics as sensitive markers of cardiovascular vulnerability.

Together, these findings establish that sleep architecture and, critically, sleep-stage dynamics encode valuable prognostic information about cardiovascular health. The ability to identify SDB from sleep patterns alone, and to predict cardiovascular outcomes with comparable accuracy even without direct respiratory measurements, demonstrates that sleep carries integrative signatures of systemic health. By capturing both monotonic trends in established risk factors and previously unreported non-linear associations reflecting physiological optima, our work extends current research and highlights the added value of flexible modelling approaches. Importantly, some of the prognostic alterations in sleep dynamics may represent downstream effects of conditions beyond SDB, such as metabolic, neuropsychiatric, or degenerative disorders, that simultaneously associate with sleep and cardiovascular outcomes. Rare transitions and disruptions of stage continuity, in particular, provide novel diagnostic and prognostic markers with potential utility for clinical risk stratification. With the increasing availability of large-scale sleep data from wearable devices and automated scoring systems, these results highlight a pathway toward unobtrusive cardiovascular risk screening and long-term monitoring. Ultimately, incorporating sleep-stage dynamics into precision medicine frameworks could enhance early detection of at-risk individuals and support preventive interventions targeting cardiovascular morbidity and mortality.

Across these three studies, we demonstrated that sleep-stage dynamics can offer robust digital biomarkers of health and disease. Using complementary approaches—causal inference in OSA, Bayesian Networks in CFS/FM, and machine learning for cardiovascular outcomes—we showed that dynamics capture disorder-specific alterations, support diagnostic differentiation, and provide prognostic information beyond traditional PSG indices. With their non-invasive nature and suitability for long-term monitoring, sleep-stage dynamics hold substantial promise for translation into home-based assessment and precision medicine. Lastly, given the predictive power of sleep in quantifying cardiovascular risk, future clinical practice may consider incorporating sleep assessment as an extension to established risk scores, such as the Framingham Risk Score [248], [249] or SCORE2 [250].

8.2 Conclusions

This dissertation explored the methodological, ethical, and translational aspects of computational sleep research across two interconnected lines of work. Specifically, it examined how automated sleep scoring (ASS) can be integrated into clinical practice to support expert decision-making, while also demonstrating how sleep-stage dynamics can be leveraged to uncover novel digital biomarkers for sleep disorders and related health conditions.

The first part (Chapters 2–4) focused on the integration of ASS into clinical practice. Here, the key contributions included (i) establishing mechanisms for effective human oversight based on uncertainty estimation, thereby aligning algorithmic predictions with physician responsibility; and (ii) developing a flexible framework to quantify predictive algorithmic bias in both clinical markers and performance metrics. These works demonstrated that fairness, transparency, and reliability, core requirements of ethical and legal mandates such as the EU AI Act and MDR, can be systematically assessed and improved. While illustrated on ASS, the proposed approaches are general in their design and applicable beyond sleep scoring, offering tools for evaluating and certifying predictive algorithms in other areas of healthcare and beyond. Together, they provide a pathway for deploying ASS systems in ways that are clinically trustworthy, interpretable, and compliant with regulatory standards.

The second part (Chapters 5–7) investigated novel digital biomarkers derived from sleep-stage dynamics, with applications spanning sleep-disordered breathing, chronic fatigue and pain syndromes, and long-term cardiovascular outcomes. Across these studies, we applied complementary methodological frameworks to address the challenge of confounding in observational or clinical data: causal meta-learners in OSA, Bayesian Networks in CFS/FM, and Random (Survival) Forests for SDB and cardiovascular risk. Whereas the first two approaches explicitly targeted causal estimation of disease effects on sleep dynamics, the latter relied on associative learning applied to prospective longitudinal data, demonstrating that the revealed associations were predictive of future outcomes. Importantly, explainability

techniques via partial effects allowed mechanistic insights into how individual predictors relate to both current sleep disorder and long-term cardiovascular vulnerability.

Taken together, our work suggests and supports sleep-stage dynamics as a class of digital markers that extend standard PSG assessments, which report basic sleep macrostructure metrics only. Beyond them, markers of sleep dynamics provide a means to assess diverse physiological signatures, distinguish between clinical conditions, and uncover prognostic information relevant for future health outcomes. With the increasing availability and widespread adoption of consumer-grade wearables, these findings highlight the potential for longitudinal applications, where sleep dynamics can be monitored unobtrusively in individuals or across clinical cohorts over extended periods. Such approaches could enable scalable, real-world screening and follow-up, ultimately contributing to precision medicine and preventive healthcare.

In summary, this dissertation pursued two complementary goals. The first was to develop strategies for integrating ASS into clinical workflows in a transparent and fair manner, providing tools for efficient human-in-the-loop review and systematic assessment of algorithmic biases. The second moved beyond sleep-stage classification to demonstrate how information embedded in polysomnographic (PSG) recordings—particularly sleep-stage dynamics—can be leveraged as digital biomarkers to support diagnosis, clinical differentiation, and risk stratification across diverse health conditions. Collectively, these contributions show that automated methods can both facilitate and standardize clinical sleep scoring, while also revealing that sleep encodes valuable physiological and predictive signals, underscoring its potential as a cornerstone of precision medicine and long-term digital health monitoring.

8.3 Limitations

This thesis is subject to several limitations that should be acknowledged. In the first branch (Chapters 2–4), although we showed that uncertainty estimation can enable efficient human oversight and introduced a flexible framework to quantify algorithmic bias, the evaluations were restricted to automated sleep scoring and a selected set of PSG-derived features. Broader application to other predictive systems in healthcare and beyond, as well as prospective testing in real-world clinical workflows, remains necessary to confirm generalizability and applicability of the proposed frameworks. Another limitation is that the proposed approaches currently lack automated open-source implementations, which restricts their immediate uptake by the wider research and clinical community and limits opportunities for independent validation and extension. First steps in this direction have already been taken, as the proposed bias-quantification framework has been incorporated into the SLEEPYLAND benchmarking platform [73], and integration of the uncertainty-based oversight methods is planned for future work.

In the second branch (Chapters 5–7), each study was subject to specific data-related and methodological constraints. Both the OSA (Chapter 5) and CFS/FM (Chapter 6) studies relied on a single PSG night per subject, which limits the ability to capture night-to-night variability in sleep dynamics. The OSA analysis was further based on observational clinical data, which—even after applying causal inference techniques—cannot fully exclude residual confounding and may reduce external validity to the general population. The CFS/FM study, while conducted on a strictly controlled experimental dataset that minimized variability and confounding, was restricted to middle-aged women, limiting its generalizability to other populations. The cardiovascular risk study (Chapter 7) benefitted from a large prospective cohort, yet its reliance on associative machine learning means that causal interpretation of the results is not warranted. In addition, individual cardiovascular outcomes (e.g., stroke, myocardial infarction) were pooled into a composite endpoint, potentially masking condition-specific patterns or the competing risks. Finally, the markers of sleep dynamics were derived from human-scored hypnograms and may therefore be affected by inter-scorer variability and annotation noise, particularly in datasets where certain scorers contributed disproportionately, although this could not be verified with the data available in our studies. Looking ahead, it may be valuable to quantify digital markers of sleep dynamics derived

also from ASS-predicted hypnograms, which could prove more robust to scorer-related variability and allow analyses at a finer temporal granularity than the conventional 30-second window.

In summary, across all studies, some broader challenges remain. Most analyses were retrospective in design and based on PSG data collected in controlled or clinical settings, which may not fully capture naturalistic sleep or long-term variability. Validation on independent datasets was conducted, but further large-scale replication, including in prospective, longitudinal, or wearable-based cohorts, is essential to establish robustness and scalability. Addressing these limitations, especially through open-source dissemination, multi-cohort validation, and integration with consumer technologies, will be key for translating the presented methods and findings into clinical and public health practice.

Bibliography

[1] B. für Statistik (BFS), *Schlafstörungen in der Bevölkerung*, DE. Neuchâtel: Bundesamt für Statistik (BFS), 2024, p. 8. [Online]. Available: https://dam-api.bfs.admin.ch/hub/api/dam/assets/32290000/master.

[2] N. M. Punjabi, B. S. Caffo, J. L. Goodwin, *et al.*, "Sleep-disordered breathing and mortality: A prospective cohort study," *PLoS medicine*, vol. 6, no. 8, e1000132, 2009.

[3] A. D. Krystal, "Psychiatric disorders and sleep," *Neurologic clinics*, vol. 30, no. 4, pp. 1389–1413, 2012.

[4] P. E. Peppard, T. Young, M. Palta, and J. Skatrud, "Prospective study of the association between sleep-disordered breathing and hypertension," *New England Journal of Medicine*, vol. 342, no. 19, pp. 1378–1384, 2000.

[5] T. Kasai, J. S. Floras, and T. D. Bradley, "Sleep apnea and cardiovascular disease: A bidirectional relationship," *Circulation*, vol. 126, no. 12, pp. 1495–1510, 2012.

[6] D. Freeman, B. Sheaves, F. Waite, A. G. Harvey, and P. J. Harrison, "Sleep disturbance and psychiatric disorders," *The Lancet Psychiatry*, vol. 7, no. 7, pp. 628–637, 2020.

[7] A. S. Gami, E. J. Olson, W. K. Shen, *et al.*, "Obstructive sleep apnea and the risk of sudden cardiac death: A longitudinal study of 10,701 adults," *Journal of the American College of Cardiology*, vol. 62, no. 7, pp. 610–616, 2013.

[8] R. B. Berry, R. Brooks, C. Gamaldo, *et al.*, *Aasm scoring manual updates for 2017 (version 2.4)*, 2017.

[9] M. I. Boulos, T. Jairam, T. Kendzerska, J. Im, A. Mekhael, and B. J. Murray, "Normal polysomnography parameters in healthy adults: A systematic review and meta-analysis," *The Lancet Respiratory Medicine*, vol. 7, no. 6, pp. 533–543, 2019.

[10] V. K. Kapur, D. H. Auckley, S. Chowdhuri, *et al.*, "Clinical practice guideline for diagnostic testing for adult obstructive sleep apnea: An american academy of sleep medicine clinical practice guideline," *Journal of clinical sleep medicine*, vol. 13, no. 3, pp. 479–504, 2017.

[11] M. H. Silber, S. Ancoli-Israel, M. H. Bonnet, *et al.*, "The visual scoring of sleep in adults," *Journal of clinical sleep medicine*, vol. 3, no. 02, pp. 121–131, 2007.

[12] L. Boswell, *Polysomnography*, Accessed: 2025-08-06, 2024. [Online]. Available: https://www.bozwell.co.uk/poly.html.

[13] A. K. Patel, V. Reddy, K. R. Shumway, and J. F. Araujo, "Physiology, sleep stages," in *StatPearls [Internet]*, StatPearls Publishing, 2024.

[14] R. S. Rosenberg and S. Van Hout, "The american academy of sleep medicine inter-scorer reliability program: Sleep stage scoring," *Journal of clinical sleep medicine*, vol. 9, no. 1, pp. 81–87, 2013.

[15] H. Danker-Hopfe, P. Anderer, J. Zeitlhofer, *et al.*, "Interrater reliability for sleep scoring according to the rechtschaffen & kales and the new aasm standard," *Journal of sleep research*, vol. 18, no. 1, pp. 74–84, 2009.

[16] A. Malhotra, M. Younes, S. T. Kuna, *et al.*, "Performance of an automated polysomnography scoring system versus computer-assisted manual scoring," *Sleep*, vol. 36, no. 4, pp. 573–582, 2013.

[17] M. M. Ohayon, M. A. Carskadon, C. Guilleminault, and M. V. Vitiello, "Meta-analysis of quantitative sleep parameters from childhood to old age in healthy individuals: Developing normative sleep values across the human lifespan," *Sleep*, vol. 27, no. 7, pp. 1255–1273, 2004.

[18] S. Redline, H. L. Kirchner, S. F. Quan, D. J. Gottlieb, V. Kapur, and A. Newman, "The effects of age, sex, ethnicity, and sleep-disordered breathing on sleep architecture," *Archives of internal medicine*, vol. 164, no. 4, pp. 406–418, 2004.

[19] E. Van Cauter, R. Leproult, and L. Plat, "Age-related changes in slow wave sleep and rem sleep and relationship with growth hormone and cortisol levels in healthy men," *Jama*, vol. 284, no. 7, pp. 861–868, 2000.

[20] E. O. Bixler, M. N. Papaliaga, A. N. Vgontzas, *et al.*, "Women sleep objectively better than men and the sleep of young women is more resilient to external stressors: Effects of age and menopause," *Journal of sleep research*, vol. 18, no. 2, pp. 221–228, 2009.

[21] J. A. Mong and D. M. Cusmano, "Sex differences in sleep: Impact of biological sex and sex steroids," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 371, no. 1688, p. 20 150 110, 2016.

[22] J. D. Edinger, M. H. Bonnet, R. R. Bootzin, *et al.*, "Derivation of research diagnostic criteria for insomnia: Report of an american academy of sleep medicine work group," *Sleep*, vol. 27, no. 8, pp. 1567–1596, 2004.

[23] J. Reiter, E. Katz, T. E. Scammell, and K. Maski, "Usefulness of a nocturnal soremp for diagnosing narcolepsy with cataplexy in a pediatric population," *Sleep*, vol. 38, no. 6, pp. 859–865, 2015.

[24] M. Billiard and K. Sonka, "Idiopathic hypersomnia," *Sleep medicine reviews*, vol. 29, pp. 23–33, 2016.

[25] R. J. Kimoff, "Sleep fragmentation in obstructive sleep apnea," *Sleep*, vol. 19, no. suppl_9, S61–S66, 1996.

[26] M. T. Bianchi, N. A. Eiseman, S. S. Cash, J. Mietus, C.-K. PENG, and R. J. Thomas, "Probabilistic sleep architecture models in patients with and without sleep apnea," *Journal of sleep research*, vol. 21, no. 3, pp. 330–341, 2012.

[27] D. Y. Goh, P. Galster, and C. L. Marcus, "Sleep architecture and respiratory disturbances in children with obstructive sleep apnea," *American journal of respiratory and critical care medicine*, vol. 162, no. 2, pp. 682–686, 2000.

[28] J. Montplaisir, J.-F. Gagnon, M. L. Fantini, *et al.*, "Polysomnographic diagnosis of idiopathic rem sleep behavior disorder," *Movement disorders*, vol. 25, no. 13, pp. 2044–2051, 2010.

[29] O. Lapierre and J. Montplaisir, "Polysomnographic features of rem sleep behavior disorder: Development of a scoring method," *Neurology*, vol. 42, no. 7, pp. 1371–1371, 1992.

[30] Y. Dauvilliers, S. Rompré, J.-F. Gagnon, M. Vendette, D. Petit, and J. Montplaisir, "Rem sleep characteristics in narcolepsy and rem sleep behavior disorder," *Sleep*, vol. 30, no. 7, pp. 844–849, 2007.

[31] A. Steiger and M. Pawlowski, "Depression and sleep," *International journal of molecular sciences*, vol. 20, no. 3, p. 607, 2019.

[32] M. BERGER and D. RIEMANN, "Rem sleep in depression—an overview," *Journal of sleep research*, vol. 2, no. 4, pp. 211–223, 1993.

[33] R. K. Malhotra, "Neurodegenerative disorders and sleep," *Sleep medicine clinics*, vol. 13, no. 1, pp. 63–70, 2018.

[34] A. McHill and K. Wright Jr, "Role of sleep and circadian disruption on energy expenditure and in metabolic predisposition to human obesity and metabolic disease," *Obesity reviews*, vol. 18, pp. 15–24, 2017.

[35] E. A. Iliescu, K. E. Yeates, and D. C. Holland, "Quality of sleep in patients with chronic kidney disease," *Nephrology Dialysis Transplantation*, vol. 19, no. 1, pp. 95–99, 2004.

[36] F. Sixel-Döring, E. Trautmann, B. Mollenhauer, and C. Trenkwalder, "Age, drugs, or disease: What alters the macrostructure of sleep in parkinson's disease?" *Sleep Medicine*, vol. 13, no. 9, pp. 1178–1183, 2012.

[37] M. M. Fung, K. Peters, S. Redline, *et al.*, "Decreased slow wave sleep increases risk of developing hypertension in elderly men," *Hypertension*, vol. 58, no. 4, pp. 596–603, 2011.

[38] S. Javaheri, Y. Y. Zhao, N. M. Punjabi, S. F. Quan, D. J. Gottlieb, and S. Redline, "Slow-wave sleep is associated with incident hypertension: The sleep heart health study," *Sleep*, vol. 41, no. 1, zsx179, 2018.

[39] E. B. Leary, K. T. Watson, S. Ancoli-Israel, *et al.*, "Association of rapid eye movement sleep with mortality in middle-aged and older adults," *JAMA neurology*, vol. 77, no. 10, pp. 1241–1251, 2020.

[40] S. Ai, S. Ye, G. Li, *et al.*, "Association of disrupted delta wave activity during sleep with long-term cardiovascular disease and mortality," *Journal of the American College of Cardiology*, vol. 83, no. 17, pp. 1671–1684, 2024.

[41] H. Khan, D. Kella, S. K. Kunutsor, K. Savonen, and J. A. Laukkanen, "Sleep duration and risk of fatal coronary heart disease, sudden cardiac death, cancer death, and all-cause mortality," *The American journal of medicine*, vol. 131, no. 12, pp. 1499–1505, 2018.

[42] C. S. Kwok, E. Kontopantelis, G. Kuligowski, *et al.*, "Self-reported sleep duration and quality and cardiovascular disease and mortality: A dose-response meta-analysis," *Journal of the American Heart Association*, vol. 7, no. 15, e008552, 2018.

[43] P. Heslop, G. D. Smith, C. Metcalfe, J. Macleod, and C. Hart, "Sleep duration and mortality: The effect of short or long sleep duration on cardiovascular and all-cause mortality in working men and women," *Sleep medicine*, vol. 3, no. 4, pp. 305–314, 2002.

[44] B.-H. Huang, M. J. Duncan, P. A. Cistulli, N. Nassar, M. Hamer, and E. Stamatakis, "Sleep and physical activity in relation to all-cause, cardiovascular disease and cancer mortality risk," *British journal of sports medicine*, vol. 56, no. 13, pp. 718–724, 2022.

[45] MDsave Inc., *Sleep study (polysomnography)*, https://www.mdsave.com/procedures/sleep-study-polysomnography/d782f4c8, Accessed: 6.8.2025, 2025.

[46] T. Taulli (GoodRx Health Editorial Team), *How much does a sleep study cost?* https://www.goodrx.com/health-topic/procedures/how-much-sleep-study-cost?, Accessed: 6.8.2025, 2022.

[47] J. Westphal (Schlafmedizinisches Zentrum Zürich), *Kosten für schlafdiagnostik und therapie*, https://sleeplab.ch/schlaftherapie-kosten/, Accessed: 6.8.2025, 2025.

[48] L. Fiorillo, A. Puiatti, M. Papandrea, *et al.*, "Automated sleep scoring: A review of the latest approaches," *Sleep medicine reviews*, vol. 48, p. 101 204, 2019.

[49] L Molinari, G Dumermuth, and B Lange, "Eeg-based multivariate statistical analysis of sleep stages," *Neuropsychobiology*, vol. 11, no. 2, pp. 140–148, 1984.

[50] H.-J. Park, J.-S. Oh, D.-U. Jeong, and K.-S. Park, "Automated sleep stage scoring using hybrid rule-and case-based reasoning," *Computers and Biomedical Research*, vol. 33, no. 5, pp. 330–349, 2000.

[51] J.-S. Oh, H.-J. Park, J.-W. Seo, and K.-S. Park, "Automatic sleep scoring based on modular rule-based reasoning units and signal processing units," in *2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and biology Society*, IEEE, vol. 2, 2001, pp. 1699–1702.

[52] S. Güneş, K. Polat, and Ş. Yosunkaya, "Efficient sleep stage recognition system based on eeg signal using k-means clustering based feature weighting," *Expert Systems with Applications*, vol. 37, no. 12, pp. 7922–7928, 2010.

[53] L. Fraiwan, K. Lweesy, N. Khasawneh, M. Fraiwan, H. Wenz, and H. Dickhaus, "Classification of sleep stages using multi-wavelet time frequency entropy and lda," *Methods of information in Medicine*, vol. 49, no. 03, pp. 230–237, 2010.

[54] L. Fraiwan, K. Lweesy, N. Khasawneh, H. Wenz, and H. Dickhaus, "Automated sleep stage identification system based on time–frequency analysis of a single eeg channel and random forest classifier," *Computer methods and programs in biomedicine*, vol. 108, no. 1, pp. 10–19, 2012.

[55] B. Şen, M. Peker, A. Çavuşoğlu, and F. V. Çelebi, "A comparative study on classification of sleep stage based on eeg signals using feature selection and classification algorithms," *Journal of medical systems*, vol. 38, no. 3, p. 18, 2014.

[56] T. Lajnef, S. Chaibi, P. Ruby, *et al.*, "Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines," *Journal of neuroscience methods*, vol. 250, pp. 94–105, 2015.

[57] L. Zoubek, S. Charbonnier, S. Lesecq, A. Buguet, and F. Chapotot, "Feature selection for sleep/wake stages classification using data driven methods," *Biomedical Signal Processing and Control*, vol. 2, no. 3, pp. 171–179, 2007.

[58] R. Vallat and M. P. Walker, "An open-source, high-performance tool for automated sleep staging," *Elife*, vol. 10, e70092, 2021.

[59] M. Perslev, S. Darkner, L. Kempfner, M. Nikolic, P. J. Jennum, and C. Igel, "U-sleep: Resilient high-frequency sleep staging," *NPJ digital medicine*, vol. 4, no. 1, p. 72, 2021.

[60] L. Fiorillo, G. Monachino, J. van der Meer, *et al.*, "U-sleep's resilience to aasm guidelines," *NPJ digital medicine*, vol. 6, no. 1, p. 33, 2023.

[61] J. B. Stephansen, A. N. Olesen, M. Olsen, *et al.*, "Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy," *Nature communications*, vol. 9, no. 1, p. 5229, 2018.

[62] A. N. Olesen, P. Jørgen Jennum, E. Mignot, and H. B. D. Sorensen, "Automatic sleep stage classification with deep residual networks in a mixed-cohort setting," *Sleep*, vol. 44, no. 1, zsaa161, 2021.

[63] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "Seqsleepnet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 3, pp. 400–410, 2019.

[64] H. Phan, K. Mikkelsen, O. Y. Chén, P. Koch, A. Mertins, and M. De Vos, "Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 8, pp. 2456–2467, 2022.

[65] T. Penzel, X. Zhang, and I. Fietze, "Inter-scorer reliability between sleep centers can teach us what to improve in the scoring rules," *Journal of Clinical Sleep Medicine*, vol. 9, no. 1, pp. 89–91, 2013.

[66] Y. J. Lee, J. Y. Lee, J. H. Cho, and J. H. Choi, "Interrater reliability of sleep stage scoring: A meta-analysis," *Journal of Clinical Sleep Medicine*, vol. 18, no. 1, pp. 193–202, 2022.

[67] H. Danker-Hopfe, D. Kunz, G. Gruber, *et al.*, "Interrater reliability between scorers from eight european sleep laboratories in subjects with different sleep disorders," *Journal of sleep research*, vol. 13, no. 1, pp. 63–69, 2004.

[68] X. Zhang, X. Dong, J. W. Kantelhardt, *et al.*, "Process and outcome for international reliability in sleep scoring," *Sleep and Breathing*, vol. 19, pp. 191–195, 2015.

[69] R. G. Norman, I. Pal, C. Stewart, J. A. Walsleben, and D. M. Rapoport, "Interobserver agreement among sleep scorers from different centers in a large dataset.," *Sleep*, vol. 23, no. 7, pp. 901–908, 2000.

[70] A. Guillot, F. Sauvet, E. H. During, and V. Thorey, "Dreem open datasets: Multi-scored sleep datasets to compare human and automated sleep staging," *IEEE transactions on neural systems and rehabilitation engineering*, vol. 28, no. 9, pp. 1955–1965, 2020.

[71] L. Fiorillo, D. Pedroncelli, V. Agostini, P. Favaro, and F. D. Faraci, "Multi-scored sleep databases: How to exploit the multiple-labels in automated sleep scoring," *Sleep*, vol. 46, no. 5, zsad028, 2023.

[72] J. P. Bakker, M. Ross, A. Cerny, *et al.*, "Scoring sleep with artificial intelligence enables quantification of sleep stage ambiguity: Hypnodensity based on multiple expert scorers and auto-scoring," *Sleep*, vol. 46, no. 2, zsac154, 2023.

[73] A. D. Rossi, M. Metaldi, M. Bechny, *et al.*, "Sleepyland: Trust begins with fair evaluation of automatic sleep staging models," *arXiv preprint arXiv:2506.08574*, 2025.

[74] G. Deng, M. Niu, S. Rao, *et al.*, "A unified flexible large polysomnography model for sleep staging and mental disorder diagnosis," *medRxiv*, pp. 2024–12, 2025.

[75] A. Kales, A. Rechtschaffen, L. A. B. I. S. University of California, and N. N. I. N. (U.S.), *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects: Allan Rechtschaffen and Anthony Kales, Editors* (NIH publication). U. S. National Institute of Neurological Diseases and Blindness, Neurological Information Network, 1968. [Online]. Available: https://books.google.ch/books?id=Z41IvQEACAAJ.

[76] D. Moser, P. Anderer, G. Gruber, *et al.*, "Sleep classification according to aasm and rechtschaffen & kales: Effects on sleep scoring parameters," *Sleep*, vol. 32, no. 2, pp. 139–149, 2009.

[77] M. M. Grigg-Damberger, "The aasm scoring manual four years later," *Journal of Clinical Sleep Medicine*, vol. 8, no. 3, pp. 323–332, 2012.

[78] W. Moraes, R. Piovezan, D. Poyares, L. R. Bittencourt, R. Santos-Silva, and S. Tufik, "Effects of aging on sleep structure throughout adulthood: A population-based study," *Sleep medicine*, vol. 15, no. 4, pp. 401–409, 2014.

[79] H Gaudreau, J Carrier, and J Montplaisir, "Age-related modifications of nrem sleep eeg: From childhood to middle age," *Journal of sleep research*, vol. 10, no. 3, pp. 165–172, 2001.

[80] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari, "Learning with noisy labels," *Advances in neural information processing systems*, vol. 26, 2013.

[81] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning from noisy labels with deep neural networks: A survey," *IEEE transactions on neural networks and learning systems*, vol. 34, no. 11, pp. 8135–8153, 2022.

[82] P. Edouard, D. Campo, P. Bartet, *et al.*, "Validation of the withings sleep analyzer, an under-the-mattress device for the detection of moderate-severe sleep apnea syndrome," *Journal of Clinical Sleep Medicine*, vol. 17, no. 6, pp. 1217–1227, 2021.

[83] J. C. Kanady, L. Ruoff, L. D. Straus, *et al.*, "Validation of sleep measurement in a multisensor consumer grade wearable device in healthy young adults," *Journal of Clinical Sleep Medicine*, vol. 16, no. 6, pp. 917–924, 2020.

[84] B. P. Choo, Y. Mok, H. C. Oh, *et al.*, "Benchmarking performance of an automatic polysomnography scoring system in a population with suspected sleep disorders," *Frontiers in Neurology*, vol. 14, p. 1 123 935, 2023.

[85] S. Gerke, T. Minssen, and G. Cohen, "Ethical and legal challenges of artificial intelligence-driven healthcare," in *Artificial intelligence in healthcare*, Elsevier, 2020, pp. 295–336.

[86] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," *Nature medicine*, vol. 25, no. 1, pp. 44–56, 2019.

[87] E. H. Shortliffe and M. J. Sepúlveda, "Clinical decision support in the era of artificial intelligence," *Jama*, vol. 320, no. 21, pp. 2199–2200, 2018.

[88] M. Younes, J. Raneri, and P. Hanly, "Staging sleep in polysomnograms: Analysis of inter-scorer variability," *Journal of Clinical Sleep Medicine*, vol. 12, no. 6, pp. 885–894, 2016.

[89] M. D. Abràmoff, M. E. Tarver, N. Loyo-Berrios, *et al.*, "Considerations for addressing bias in artificial intelligence for health equity," *NPJ digital medicine*, vol. 6, no. 1, p. 170, 2023.

[90] J. W. Gichoya, K. Thomas, L. A. Celi, *et al.*, "Ai pitfalls and what not to do: Mitigating bias in ai," *The British Journal of Radiology*, vol. 96, no. 1150, p. 20 230 023, 2023.

[91] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM computing surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[92] N. Norori, Q. Hu, F. M. Aellen, F. D. Faraci, and A. Tzovara, "Addressing bias in big data and ai for health care: A call for open science," *Patterns*, vol. 2, no. 10, 2021.

[93] H. van Gorp, I. A. Huijben, P. Fonseca, R. J. van Sloun, S. Overeem, and M. M. van Gilst, "Certainty about uncertainty in sleep staging: A theoretical framework," *Sleep*, vol. 45, no. 8, zsac134, 2022.

[94] J. K. Hong, T. Lee, R. D. Delos Reyes, *et al.*, "Confidence-based framework using deep learning for automated sleep stage scoring," *Nature and Science of Sleep*, pp. 2239–2250, 2021.

[95] M. Bechny, G. Monachino, L. Fiorillo, *et al.*, "Bridging ai and clinical practice: Integrating automated sleep scoring algorithm with uncertainty-guided physician review," *Nature and science of sleep*, pp. 555–572, 2024.

[96] M. Bechny, G. Monachino, L. Fiorillo, *et al.*, "Framework for algorithmic bias quantification and its application to automated sleep scoring," in *2024 11th IEEE Swiss Conference on Data Science (SDS)*, IEEE, 2024, pp. 250–253.

[97] M. Bechny, L. Fiorillo, J. van der Meer, *et al.*, "Beyond accuracy: A framework for evaluating algorithmic bias and performance, applied to automated sleep scoring," *Scientific Reports*, vol. 15, no. 1, p. 21 421, 2025.

[98] B. Kemp and H. A. Kamphuisen, "Simulation of human hypnograms using a markov chain model," *Sleep*, vol. 9, no. 3, pp. 405–414, 1986.

[99] A. Yassouridis, A. Steiger, A. Klinger, and L. Fahrmeir, "Modelling and exploring human sleep with event history analysis," *Journal of sleep research*, vol. 8, no. 1, pp. 25–36, 1999.

[100] J. W. Burns, L. J. Crofford, and R. D. Chervin, "Sleep stage dynamics in fibromyalgia patients and controls," *Sleep Medicine*, vol. 9, no. 6, pp. 689–696, 2008.

[101] A. Laffan, B. Caffo, B. J. Swihart, and N. M. Punjabi, "Utility of sleep stage transitions in assessing sleep continuity," *Sleep*, vol. 33, no. 12, pp. 1681–1686, 2010.

[102] A. Kishi, Z. R. Struzik, B. H. Natelson, F. Togo, and Y. Yamamoto, "Dynamics of sleep stage transitions in healthy humans and patients with chronic fatigue syndrome," *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, vol. 294, no. 6, R1980–R1987, 2008.

[103] J. Kim, J.-S. Lee, P. Robinson, and D.-U. Jeong, "Markov analysis of sleep dynamics," *Physical review letters*, vol. 102, no. 17, p. 178 104, 2009.

[104] Y. Wei, M. A. Colombo, J. R. Ramautar, *et al.*, "Sleep stage transition dynamics reveal specific stage 2 vulnerability in insomnia," *Sleep*, vol. 40, no. 9, zsx117, 2017.

[105] A. Schlemmer, U. Parlitz, S. Luther, N Wessel, and T. Penzel, "Changes of sleep-stage transitions due to ageing and sleep disorder," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 373, no. 2034, p. 20 140 093, 2015.

[106] M. Wächter, J. W. Kantelhardt, M. R. Bonsignore, *et al.*, "Unique sleep-stage transitions determined by obstructive sleep apnea severity, age and gender," *Journal of sleep research*, vol. 29, no. 2, e12895, 2020.

[107] B. D. Yetton, E. A. McDevitt, N. Cellini, C. Shelton, and S. C. Mednick, "Quantifying sleep architecture dynamics and individual differences using big data and bayesian networks," *PloS one*, vol. 13, no. 4, e0194604, 2018.

[108] C.-C. Lo, L. N. Amaral, S. Havlin, *et al.*, "Dynamics of sleep-wake transitions during sleep," *Europhysics Letters*, vol. 57, no. 5, p. 625, 2002.

[109] T. Penzel, J. W. Kantelhardt, C.-C. Lo, K. Voigt, and C. Vogelmeier, "Dynamics of heart rate and sleep stages in normals and patients with sleep apnea," *Neuropsychopharmacology*, vol. 28, no. 1, S48–S53, 2003.

[110] R. G. Norman, M. A. Scott, I. Ayappa, J. A. Walsleben, and D. M. Rapoport, "Sleep continuity measured by survival curve analysis," *Sleep*, vol. 29, no. 12, pp. 1625–1631, 2006.

[111] R. D. Chervin, J. L. Fetterolf, D. L. Ruzicka, B. J. Thelen, and J. W. Burns, "Sleep stage dynamics differ between children with and without obstructive sleep apnea," *Sleep*, vol. 32, no. 10, pp. 1325–1332, 2009.

[112] M. T. Bianchi, S. S. Cash, J. Mietus, C.-K. Peng, and R. Thomas, "Obstructive sleep apnea alters sleep stage transition dynamics," *PLoS One*, vol. 5, no. 6, e11356, 2010.

[113] E. B. Klerman, W. Wang, J. F. Duffy, D.-J. Dijk, C. A. Czeisler, and R. E. Kronauer, "Survival analysis indicates that age-related decline in sleep continuity occurs exclusively during nrem sleep," *Neurobiology of aging*, vol. 34, no. 1, pp. 309–318, 2013.

[114] A. Kishi, S. Haraki, R. Toyota, *et al.*, "Sleep stage dynamics in young patients with sleep bruxism," *Sleep*, vol. 43, no. 1, zsz202, 2020.

[115] M. Bechny, A. Kishi, L. Fiorillo, *et al.*, "Novel digital markers of sleep dynamics: Causal inference approach revealing age and gender phenotypes in obstructive sleep apnea," *Scientific Reports*, vol. 15, no. 1, p. 12 016, 2025.

[116] M. Bechny, M. Scutari, J. van der Meer, *et al.*, "Unveiling sleep dysregulation in chronic fatigue syndrome with and without fibromyalgia through bayesian networks," in *International Conference on Artificial Intelligence in Medicine*, Springer, 2025, pp. 33–43.

[117] M. Bechny, A. Kishi, Y. Tomita, *et al.*, "Sleep-stage dynamics predict current sleep-disordered breathing and future cardiovascular risk," *medRxiv*, pp. 2025–07, 2025, Preprint.

[118] G. Dorffner, M. Vitr, and P. Anderer, "The effects of aging on sleep architecture in healthy subjects," in *GeNeDis 2014: Geriatrics*, Springer, 2014, pp. 93–100.

[119] A. Kahn, B. Dan, J. Groswasser, P. Franco, and M Sottiaux, "Normal sleep architecture in infants and children," *Journal of Clinical Neurophysiology*, vol. 13, no. 3, pp. 184–197, 1996.

[120] Y. Liu, J. Zhang, V. Lam, *et al.*, "Altered sleep stage transitions of rem sleep: A novel and stable biomarker of narcolepsy," *Journal of Clinical Sleep Medicine*, vol. 11, no. 8, pp. 885–894, 2015.

[121] T. Penzel, *Sleep scoring moving from visual scoring towards automated scoring*, 2022.

[122] D. Y. Kang, P. N. DeYoung, J. Tantiongloc, T. P. Coleman, and R. L. Owens, "Statistical uncertainty quantification to augment clinical decision support: A first implementation in sleep medicine," *NPJ digital medicine*, vol. 4, no. 1, p. 142, 2021.

[123] L. Fiorillo, P. Favaro, and F. D. Faraci, "Deepsleepnet-lite: A simplified automatic sleep stage scoring model with uncertainty estimates," *IEEE transactions on neural systems and rehabilitation engineering*, vol. 29, pp. 2076–2085, 2021.

[124] M. Rusanen, G. Jouan, R. Huttunen, *et al.*, "Asaga: Automatic sleep analysis with gray areas," *arXiv preprint arXiv:2310.02032*, 2023.

[125] F. M. Aellen, J. Van der Meer, A. Dietmann, M. Schmidt, C. L. Bassetti, and A. Tzovara, "Disentangling the complex landscape of sleep–wake disorders with data-driven phenotyping: A study of the bernese center," *European journal of neurology*, vol. 31, no. 1, e16026, 2024.

[126] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, Springer, 2015, pp. 234–241.

[127] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 427–436.

[128] C. Corbiere, N. Thome, A. Saporta, T.-H. Vu, M. Cord, and P. Perez, "Confidence estimation via auxiliary models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6043–6055, 2021.

[129] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[130] J. Aitchison, "The statistical analysis of compositional data," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 44, no. 2, pp. 139–160, 1982.

[131] B. Efron, *The jackknife, the bootstrap and other resampling plans*. SIAM, 1982.

[132] T. Penzel and R. Conradt, "Computer based sleep recording and analysis," *Sleep medicine reviews*, vol. 4, no. 2, pp. 131–148, 2000.

[133] S. Yalamanchali, V. Farajian, C. Hamilton, T. R. Pott, C. G. Samuelson, and M. Friedman, "Diagnosis of obstructive sleep apnea by peripheral arterial tonometry: Meta-analysis," *JAMA Otolaryngology–Head & Neck Surgery*, vol. 139, no. 12, pp. 1343–1350, 2013.

[134] J. M. Bland and D. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *The lancet*, vol. 327, no. 8476, pp. 307–310, 1986.

[135] L. Xu, F. Han, B. T. Keenan, *et al.*, "Validation of the nox-t3 portable monitor for diagnosis of obstructive sleep apnea in chinese adults," *Journal of Clinical Sleep Medicine*, vol. 13, no. 5, pp. 675–683, 2017.

[136] J. Ludbrook, "Confidence in altman–bland plots: A critical review of the method of differences," *Clinical and Experimental Pharmacology and Physiology*, vol. 37, no. 2, pp. 143–149, 2010.

[137] G. Casella and R. Berger, *Statistical inference*. Chapman and Hall/CRC, 2024.

[138] M. D. Stasinopoulos, R. A. Rigby, G. Z. Heller, V. Voudouris, and F. De Bastiani, *Flexible regression and smoothing: using GAMLSS in R*. CRC Press, Taylor & Francis Group, 2017.

[139] C. Sahlin, K. A. Franklin, H. Stenlund, and E. Lindberg, "Sleep in women: Normal values for sleep stages and position and the effect of age, obesity, sleep apnea, smoking, alcohol and hypertension," *Sleep medicine*, vol. 10, no. 9, pp. 1025–1030, 2009.

[140] S. Nikkonen, P. Somaskandhan, H. Korkalainen, *et al.*, "Multicentre sleep-stage scoring agreement in the sleep revolution project," *Journal of Sleep Research*, vol. 33, no. 1, e13956, 2024.

[141] B. Frénay and M. Verleysen, "Classification in the presence of label noise: A survey," *IEEE transactions on neural networks and learning systems*, vol. 25, no. 5, pp. 845–869, 2013.

[142] M. L. Gordon, K. Zhou, K. Patel, T. Hashimoto, and M. S. Bernstein, "The disagreement deconvolution: Bringing machine learning performance metrics in line with reality," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–14.

[143] M. Nagendran, Y. Chen, C. A. Lovejoy, *et al.*, "Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies," *bmj*, vol. 368, 2020.

[144] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods," *Machine learning*, vol. 110, no. 3, pp. 457–506, 2021.

[145] A. Goyal and Y. Bengio, "Inductive biases for deep learning of higher-level cognition," *Proceedings of the Royal Society A*, vol. 478, no. 2266, p. 20 210 068, 2022.

[146] R. Bin Rafiq, F. Modave, S. Guha, and M. V. Albert, "Validation methods to promote real-world applicability of machine learning in medicine," in *Proceedings of the 2020 3rd International Conference on Digital Medicine and Image Processing*, 2020, pp. 13–19.

[147] D. Borsboom, G. J. Mellenbergh, and J. Van Heerden, "The concept of validity.," *Psychological review*, vol. 111, no. 4, p. 1061, 2004.

[148] R. A. Rigby, M. D. Stasinopoulos, G. Z. Heller, and F. De Bastiani, *Distributions for modeling location, scale, and shape: Using GAMLSS in R*. Chapman and Hall/CRC, 2019.

[149] C. V. Senaratna, J. L. Perret, C. J. Lodge, *et al.*, "Prevalence of obstructive sleep apnea in the general population: A systematic review," *Sleep medicine reviews*, vol. 34, pp. 70–81, 2017.

[150] X. Wang, Y. Ouyang, Z. Wang, G. Zhao, L. Liu, and Y. Bi, "Obstructive sleep apnea and risk of cardiovascular disease and all-cause mortality: A meta-analysis of prospective cohort studies," *International journal of cardiology*, vol. 169, no. 3, pp. 207–214, 2013.

[151] V. Mancebo-Sosa, V. Mancilla-Hernández, J. Miranda-Ortiz, *et al.*, "Sleep architecture alterations in patients with periodic limb movements disorder during sleep and sleep breathing disorders," *Sleep Science*, vol. 9, no. 2, pp. 84–88, 2016.

[152] P. Drakatos, M. Olaithe, D. Verma, *et al.*, "Periodic limb movements during sleep: A narrative review," *Journal of Thoracic Disease*, vol. 13, no. 11, p. 6476, 2021.

[153] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2025. [Online]. Available: https://www.R-project.org/.

[154] R. Budhiraja, S. Javaheri, M. K. Pavlova, L. J. Epstein, O. Omobomi, and S. F. Quan, "Prevalence and correlates of periodic limb movements in osa and the effect of cpap therapy," *Neurology*, vol. 94, no. 17, e1820–e1827, 2020.

[155] P. H. Eilers, B. D. Marx, and M. Durbán, "Twenty years of p-splines," *SORT: statistics and operations research transactions*, vol. 39, no. 2, pp. 0149–186, 2015.

[156] J. R. D. Espiritu, "Aging-related sleep changes," *Clinics in geriatric medicine*, vol. 24, no. 1, pp. 1–14, 2008.

[157] B. A. Mander, J. R. Winer, and M. P. Walker, "Sleep and human aging," *Neuron*, vol. 94, no. 1, pp. 19–36, 2017.

[158] S. Redline, K. Kump, P. V. Tishler, I. Browner, and V. Ferrette, "Gender differences in sleep disordered breathing in a community-based sample.," *American journal of respiratory and critical care medicine*, vol. 149, no. 3, pp. 722–726, 1994.

[159] J. Haba-Rubio, H. Marti-Soler, P. Marques-Vidal, *et al.*, "Prevalence and determinants of periodic limb movements in the general population," *Annals of neurology*, vol. 79, no. 3, pp. 464–474, 2016.

[160] M. Papenberg and G. W. Klau, "Using anticlustering to partition data sets into equivalent parts.," *Psychological Methods*, vol. 26, no. 2, p. 161, 2021.

[161] P. K. Lohia, K. N. Ramamurthy, M. Bhide, D. Saha, K. R. Varshney, and R. Puri, "Bias mitigation post-processing for individual and group fairness," in *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)*, IEEE, 2019, pp. 2847–2851.

[162] M. P. Kim, A. Ghorbani, and J. Zou, "Multiaccuracy: Black-box post-processing for fairness in classification," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 247–254.

[163] S. Bird, M. Dudík, R. Edgar, *et al.*, "Fairlearn: A toolkit for assessing and improving fairness in ai," *Microsoft, Tech. Rep. MSR-TR-2020-32*, 2020.

[164] K. T. Hufthammer, T. H. Aasheim, S. Ånneland, H. Brynjulfsen, and M. Slavkovik, "Bias mitigation with aif360: A comparative study," in *NIKT: Norsk IKT-konferanse for forskning og utdanning 2020*, Norsk IKT-konferanse for forskning og utdanning, 2020.

[165] M. A. Carskadon, W. C. Dement, *et al.*, "Normal human sleep: An overview," *Principles and practice of sleep medicine*, vol. 4, no. 1, pp. 13–23, 2005.

[166] G. Luca, J. Haba Rubio, D. Andries, *et al.*, "Age and gender variations of sleep in subjects without sleep disorders," *Annals of medicine*, vol. 47, no. 6, pp. 482–491, 2015.

[167] M. Egger, M. Schneider, and G. D. Smith, "Meta-analysis spurious precision? meta-analysis of observational studies," *Bmj*, vol. 316, no. 7125, pp. 140–144, 1998.

[168] W. G. Cochran and D. B. Rubin, "Controlling bias in observational studies: A review," *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 417–446, 1973.

[169] T Penzel, C.-C. Lo, P. Ivanov, K Kesper, H. Becker, and C Vogelmeier, "Analysis of sleep fragmentation and sleep structure in patients with sleep apnea and normal volunteers," in *2005 IEEE Engineering in medicine and biology 27th annual conference*, IEEE, 2006, pp. 2591–2594.

[170] O. Andlauer, H. Moore, L. Jouhier, *et al.*, "Nocturnal rapid eye movement sleep latency for identifying patients with narcolepsy/hypocretin deficiency," *JAMA neurology*, vol. 70, no. 7, pp. 891–902, 2013.

[171] L. W. Hermans, I. A. Huijben, H. van Gorp, *et al.*, "Representations of temporal sleep dynamics: Review and synthesis of the literature," *Sleep Medicine Reviews*, vol. 63, p. 101 611, 2022.

[172] C. Jackson, "Multi-state models for panel data: The msm package for r," *Journal of statistical software*, vol. 38, pp. 1–28, 2011.

[173] J. Kalbfleisch and J. F. Lawless, "The analysis of panel data under a markov assumption," *Journal of the american statistical association*, vol. 80, no. 392, pp. 863–871, 1985.

[174] J. H. Ellenberg, "Selection bias in observational and experimental studies," *Statistics in medicine*, vol. 13, no. 5-7, pp. 557–567, 1994.

[175] D. B. Rubin, "The use of matched sampling and regression adjustment to remove bias in observational studies," *Biometrics*, pp. 185–203, 1973.

[176] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[177] K. Hirano and G. W. Imbens, "Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization," *Health Services and Outcomes research methodology*, vol. 2, pp. 259–278, 2001.

[178] N. C. Chesnaye, V. S. Stel, G. Tripepi, *et al.*, "An introduction to inverse probability of treatment weighting in observational research," *Clinical Kidney Journal*, vol. 15, no. 1, pp. 14–20, 2022.

[179] M. J. Maier, *Dirichletreg: Dirichlet regression*, R package version 0.7-1, 2021. [Online]. Available: https://github.com/maiermarco/DirichletReg.

[180] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu, "Metalearners for estimating heterogeneous treatment effects using machine learning," *Proceedings of the national academy of sciences*, vol. 116, no. 10, pp. 4156–4165, 2019.

[181] G. W. Imbens and D. B. Rubin, *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

[182] S. F. Quan, B. V. Howard, C. Iber, *et al.*, "The sleep heart health study: Design, rationale, and methods," *Sleep*, vol. 20, no. 12, pp. 1077–1085, 1997.

[183] G.-Q. Zhang, L. Cui, R. Mueller, *et al.*, "The national sleep research resource: Towards a sleep data commons," *Journal of the American Medical Informatics Association*, vol. 25, no. 10, pp. 1351–1358, 2018.

[184] P. C. Austin, "Variance estimation when using inverse probability of treatment weighting (iptw) with survival analysis," *Statistics in medicine*, vol. 35, no. 30, pp. 5642–5655, 2016.

[185] C. Lal, C. Strange, and D. Bachman, "Neurocognitive impairment in obstructive sleep apnea," *Chest*, vol. 141, no. 6, pp. 1601–1610, 2012.

[186] R. Stickgold and M. P. Walker, "Sleep-dependent memory consolidation and reconsolidation," *Sleep medicine*, vol. 8, no. 4, pp. 331–343, 2007.

[187] G. Plazzi and F. Pizza, "Sleep dynamics beyond traditional sleep macrostructure," *Sleep*, vol. 36, no. 8, pp. 1123–1124, 2013.

[188] A. M. Sweetman, L. C. Lack, P. G. Catcheside, *et al.*, "Developing a successful treatment for co-morbid insomnia and sleep apnoea," *Sleep medicine reviews*, vol. 33, pp. 28–38, 2017.

[189] J. W. Winkelman, E. Shahar, I. Sharief, and D. J. Gottlieb, "Association of restless legs syndrome and cardiovascular disease in the sleep heart health study," *Neurology*, vol. 70, no. 1, pp. 35–42, 2008.

[190] A. Benetó, E. Gomez-Siurana, and P. Rubio-Sanchez, "Comorbidity between sleep apnea and insomnia," *Sleep medicine reviews*, vol. 13, no. 4, pp. 287–293, 2009.

[191] F. S. Luyster, D. J. Buysse, and P. J. Strollo Jr, "Comorbid insomnia and obstructive sleep apnea: Challenges for clinical practice and research," *Journal of Clinical Sleep Medicine*, vol. 6, no. 2, pp. 196–204, 2010.

[192] B. B. Koo, S. R. Patel, K. Strohl, and V. Hoffstein, "Rapid eye movement-related sleep-disordered breathing: Influence of age and gender," *Chest*, vol. 134, no. 6, pp. 1156–1161, 2008.

[193] M. R. Shepertycky, K. Banno, and M. H. Kryger, "Differences between men and women in the clinical presentation of patients diagnosed with obstructive sleep apnea syndrome," *Sleep*, vol. 28, no. 3, pp. 309–314, 2005.

[194] A. Valipour, H. Lothaller, H. Rauscher, H. Zwick, O. C. Burghuber, and P. Lavie, "Gender-related differences in symptoms of patients with suspected breathing disorders in sleep: A clinical population study using the sleep disorders questionnaire," *Sleep*, vol. 30, no. 3, pp. 312–319, 2007.

[195] M. Bublitz, N. Adra, L. Hijazi, F. Shaib, H. Attarian, and G. Bourjeily, "A narrative review of sex and gender differences in sleep disordered breathing: Gaps and opportunities," *Life*, vol. 12, no. 12, p. 2003, 2022.

[196] M. G. Terzano, L. Parrino, M. Boselli, M. C. Spaggiari, and G. Di Giovanni, "Polysomnographic analysis of arousal responses in obstructive sleep apnea syndrome by means of the cyclic alternating pattern," *Journal of Clinical Neurophysiology*, vol. 13, no. 2, pp. 145–155, 1996.

[197] L. Parrino, A Smerieri, M Boselli, M. Spaggiari, and M. G. Terzano, "Sleep reactivity during acute nasal cpap in obstructive sleep apnea syndrome," *Neurology*, vol. 54, no. 8, pp. 1633–1640, 2000.

[198] L. Parrino, R. J. Thomas, A. Smerieri, M. C. Spaggiari, A. Del Felice, and M. G. Terzano, "Reorganization of sleep patterns in severe osas under prolonged cpap treatment," *Clinical neurophysiology*, vol. 116, no. 9, pp. 2228–2239, 2005.

[199] G. Milioli, M. Bosi, A. Grassi, *et al.*, "Can sleep microstructure improve diagnosis of osas? integrative information from cap parameters," *Archives Italiennes de Biologie*, vol. 153, no. 2-3, pp. 194–203, 2015.

[200] C. Mutti, I. Pollara, A. Abramo, *et al.*, "The contribution of sleep texture in the characterization of sleep apnea," *Diagnostics*, vol. 13, no. 13, p. 2217, 2023.

[201] A. Kishi, B. H. Natelson, F. Togo, Z. R. Struzik, D. M. Rapoport, and Y. Yamamoto, "Sleep-stage dynamics in patients with chronic fatigue syndrome with or without fibromyalgia," *Sleep*, vol. 34, no. 11, pp. 1551–1560, 2011.

[202] M. M. Brown and L. A. Jason, "Functioning in individuals with chronic fatigue syndrome: Increased impairment with co-occurring multiple chemical sensitivity and fibromyalgia," *Dyn Med*, vol. 6, pp. 1–9, 2007.

[203] E. W. Clayton, "Beyond myalgic encephalomyelitis/chronic fatigue syndrome: An iom report on redefining an illness," *JAMA*, vol. 313, no. 11, pp. 1101–1102, 2015.

[204] C. M. Galvez-Sánchez and G. A. Reyes del Paso, "Diagnostic criteria for fibromyalgia: Critical review and future perspectives," *J Clin Med*, vol. 9, no. 4, p. 1219, 2020.

[205] M. Faro, N. Sàez-Francàs, J. Castro-Marrero, L. Aliste, T. F. de Sevilla, and J. Alegre, "Gender differences in chronic fatigue syndrome," *Reumatol Clin (Engl Ed)*, vol. 12, no. 2, pp. 72–77, 2016.

[206] B. H. Natelson, "Myalgic encephalomyelitis/chronic fatigue syndrome and fibromyalgia: Definitions, similarities, and differences," *Clin Ther*, vol. 41, no. 4, pp. 612–618, 2019.

[207] E. R. Unger, R. Nisenbaum, H. Moldofsky, *et al.*, "Sleep assessment in a population-based study of chronic fatigue syndrome," *BMC Neurol*, vol. 4, pp. 1–9, 2004.

[208] A.-S. N. Mariman, D. P. Vogelaers, E. Tobback, L. M. Delesie, I. P. Hanoulle, and D. A. Pevernagie, "Sleep in the chronic fatigue syndrome," *Sleep Med Rev*, vol. 17, no. 3, pp. 193–199, 2013.

[209] M. Scutari and J.-B. Denis, *Bayesian networks: with examples in R*, 2nd. Chapman and Hall/CRC, 2021.

[210] J. Pearl, *Causality*. Cambridge University Press, 2009.

[211] E. Bareinboim and J. Pearl, "Causal inference and the data-fusion problem," *Proceedings of the National Academy of Sciences*, vol. 113, no. 27, pp. 7345–7352, 2016.

[212] S. Redline, G. Yenokyan, D. J. Gottlieb, *et al.*, "Obstructive sleep apnea–hypopnea and incident stroke: The sleep heart health study," *American journal of respiratory and critical care medicine*, vol. 182, no. 2, pp. 269–277, 2010.

[213] L. F. Drager, R. D. McEvoy, F. Barbe, G. Lorenzi-Filho, and S. Redline, "Sleep apnea and cardiovascular disease: Lessons from recent trials and need for team science," *Circulation*, vol. 136, no. 19, pp. 1840–1850, 2017.

[214] S. Javaheri, F. Barbe, F. Campos-Rodriguez, *et al.*, "Sleep apnea: Types, mechanisms, and clinical cardiovascular consequences," *Journal of the American College of Cardiology*, vol. 69, no. 7, pp. 841–858, 2017.

[215] H. K. Yaggi, J. Concato, W. N. Kernan, J. H. Lichtman, L. M. Brass, and V. Mohsenin, "Obstructive sleep apnea as a risk factor for stroke and death," *New England Journal of Medicine*, vol. 353, no. 19, pp. 2034–2041, 2005.

[216] R. S. Leung and T Douglas Bradley, "Sleep apnea and cardiovascular disease," *American journal of respiratory and critical care medicine*, vol. 164, no. 12, pp. 2147–2165, 2001.

[217] A. S. Gami, D. E. Howard, E. J. Olson, and V. K. Somers, "Day–night pattern of sudden death in obstructive sleep apnea," *New England Journal of Medicine*, vol. 352, no. 12, pp. 1206–1214, 2005.

[218] A. N. Vgontzas, T. L. Tan, E. O. Bixler, L. F. Martin, D. Shubert, and A. Kales, "Sleep apnea and sleep disruption in obese patients," *Archives of internal medicine*, vol. 154, no. 15, pp. 1705–1711, 1994.

[219] K. Shahveisi, A. Jalali, M. R. Moloudi, S. Moradi, A. Maroufi, and H. Khazaie, "Sleep architecture in patients with primary snoring and obstructive sleep apnea," *Basic and Clinical Neuroscience*, vol. 9, no. 2, p. 147, 2018.

[220] D.-J. Dijk, "Regulation and functional correlates of slow wave sleep," *Journal of Clinical Sleep Medicine*, vol. 5, no. 2 suppl, S6–S15, 2009.

[221] J. D. Payne and L. Nadel, "Sleep, dreams, and memory consolidation: The role of the stress hormone cortisol," *Learning & Memory*, vol. 11, no. 6, pp. 671–678, 2004.

[222] M. Mendoza-Alvarez, Y. Balthasar, J. Verbraecken, *et al.*, "Systematic review: Rem sleep, dysphoric dreams and nightmares as transdiagnostic features of psychiatric disorders with emotion dysregulation-clinical implications," *Sleep Medicine*, vol. 127, pp. 1–15, 2025.

[223] R. Boyce, S. Williams, and A. Adamantidis, "Rem sleep and memory," *Current Opinion in Neurobiology*, vol. 44, pp. 167–177, 2017.

[224] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[225] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, 2016.

[226] S. J. Rigatti, "Random forest," *Journal of Insurance Medicine*, vol. 47, no. 1, pp. 31–39, 2017.

[227] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *The Annals of Applied Statistics*, vol. 2, no. 3, pp. 841 –860, 2008. DOI: 10.1214/08-AOAS169. [Online]. Available: https://doi.org/10.1214/08-AOAS169.

[228] H. Ishwaran, M. S. Lauer, E. H. Blackstone, M. Lu, and U. B. Kogalur, *randomForestSRC: Random survival forests vignette*, http://randomforestsrc.org/articles/survival.html, [accessed date], 2021. [Online]. Available: http://randomforestsrc.org/articles/survival.html.

[229] M. Papenberg and G. W. Klau, "Using anticlustering to partition data sets into equivalent parts.," *Psychological Methods*, vol. 26, no. 2, p. 161, 2021.

[230] P. E. Peppard, T. Young, J. H. Barnet, M. Palta, E. W. Hagen, and K. M. Hla, "Increased prevalence of sleep-disordered breathing in adults," *American journal of epidemiology*, vol. 177, no. 9, pp. 1006–1014, 2013.

[231] R. Heinzer, S Vat, P. Marques-Vidal, *et al.*, "Prevalence of sleep-disordered breathing in the general population: The hypnolaus study," *The Lancet Respiratory Medicine*, vol. 3, no. 4, pp. 310–318, 2015.

[232] C Filozof, M. Fernandez Pinilla, and A Fernández-Cruz, "Smoking cessation and weight gain," *Obesity reviews*, vol. 5, no. 2, pp. 95–103, 2004.

[233] K. A. Perkins, "Weight gain following smoking cessation.," *Journal of consulting and clinical psychology*, vol. 61, no. 5, p. 768, 1993.

[234] H. Ishwaran, M. Lu, and U. B. Kogalur, *randomForestSRC: Getting started with randomForestSRC vignette*, `http://randomforestsrc.org/articles/getstarted.html`, 2021. [Online]. Available: `http://randomforestsrc.org/articles/getstarted.html`.

[235] S. P. Khot and L. B. Morgenstern, "Sleep and stroke," *Stroke*, vol. 50, no. 6, pp. 1612–1617, 2019.

[236] E. Hale, E. Gottlieb, J. Usseglio, and A. Shechter, "Post-stroke sleep disturbance and recurrent cardiovascular and cerebrovascular events: A systematic review and meta-analysis," *Sleep medicine*, vol. 104, pp. 29–41, 2023.

[237] D. M. Hermann and C. L. Bassetti, "Role of sleep-disordered breathing and sleep-wake disturbances for stroke and stroke recovery," *Neurology*, vol. 87, no. 13, pp. 1407–1416, 2016.

[238] O. Ludka, R. Stepanova, M. Vyskocilova, *et al.*, "Sleep apnea prevalence in acute myocardial infarction—the sleep apnea in post-acute myocardial infarction patients (sapami) study," *International journal of cardiology*, vol. 176, no. 1, pp. 13–19, 2014.

[239] E. O. Bixler, A. N. Vgontzas, H.-M. Lin, *et al.*, "Prevalence of sleep-disordered breathing in women: Effects of gender," *American journal of respiratory and critical care medicine*, vol. 163, no. 3, pp. 608–613, 2001.

[240] K. Spiegelhalder, W. Regen, S. Nanovska, C. Baglioni, and D. Riemann, "Comorbid sleep disorders in neuropsychiatric disorders across the life cycle," *Current psychiatry reports*, vol. 15, no. 6, p. 364, 2013.

[241] S. Rao Kondapally Seshasai, S. Kaptoge, A. Thompson, *et al.*, "Diabetes mellitus, fasting glucose, and risk of cause-specific death.," *The New England journal of medicine*, vol. 364, no. 9, pp. 829–841, 2011.

[242] A. S. Go, G. M. Chertow, D. Fan, C. E. McCulloch, and C.-y. Hsu, "Chronic kidney disease and the risks of death, cardiovascular events, and hospitalization," *New England Journal of Medicine*, vol. 351, no. 13, pp. 1296–1305, 2004.

[243] A. Saeed, O. Lopez, A. Cohen, and S. E. Reis, "Cardiovascular disease and alzheimer's disease: The heart–brain axis," *Journal of the American Heart Association*, vol. 12, no. 21, e030780, 2023.

[244] A. Fayaz, S. Ayis, S. S. Panesar, R. M. Langford, and L. J. Donaldson, "Assessing the relationship between chronic pain and cardiovasculardisease: A systematic review and meta-analysis," *Scandinavian journal of pain*, vol. 13, no. 1, pp. 76–90, 2016.

[245] P. P. Ujma and R. Bódizs, "Sleep alterations as a function of 88 health indicators," *BMC medicine*, vol. 22, no. 1, p. 134, 2024.

[246] J Perrier, M Duivon, P Clochon, *et al.*, "Sleep macro-and microstructure in breast cancer survivors," *Scientific Reports*, vol. 12, no. 1, p. 2557, 2022.

[247] S. H. Armenian, L. Xu, B. Ky, *et al.*, "Cardiovascular disease among survivors of adult-onset cancer: A community-based retrospective cohort study," *Journal of Clinical Oncology*, vol. 34, no. 10, pp. 1122–1130, 2016.

[248] R. B. Schnabel, L. M. Sullivan, D. Levy, *et al.*, "Development of a risk score for atrial fibrillation (framingham heart study): A community-based cohort study," *The Lancet*, vol. 373, no. 9665, pp. 739–745, 2009.

[249] D. M. Lloyd-Jones, P. W. Wilson, M. G. Larson, *et al.*, "Framingham risk score and prediction of lifetime risk for coronary heart disease," *The American journal of cardiology*, vol. 94, no. 1, pp. 20–24, 2004.

[250] SCORE2 working group and ESC Cardiovascular Risk Collaboration, "Score2 risk prediction algorithms: New models to estimate 10-year risk of cardiovascular disease in europe," *European heart journal*, vol. 42, no. 25, pp. 2439–2454, 2021.

Appendix A

Supplementary Materials for Chapter 4

A.1 Statistical characteristics of derived PSG markers

Table A.1: Descriptive statistics of sleep metrics from physician scoring and predictions by U-Sleep and YASA.

Metric	Sleep Scoring	Mean	SD	Q10	Q25	Q50	Q75	Q90	Min	Max
Sleep Latency [*minutes*]	-	17.8	23.9	2.0	4.5	10.5	21.5	40.5	0.0	339.5
	U-Sleep	15.4	20.8	1.5	4.0	8.5	18.0	36.0	0.0	204.5
	YASA	26.3	29.2	4.0	8.5	17.0	33.0	60.0	0.0	306.5
REM Latency [*minutes*]	-	139.5	80.9	59.0	77.5	119.0	184.5	258.9	0.0	502.0
	U-Sleep	131.1	84.7	48.0	72.5	111.0	177.0	252.0	0.0	896.5
	YASA	112.4	66.4	45.2	67.5	97.5	149.5	201.5	0.0	486.5
Total Sleep Time [*minutes*]	-	338.3	89.0	239.5	291.5	337.5	382.0	424.8	0.0	848.5
	U-Sleep	343.6	89.4	247.0	297.0	342.0	385.0	428.5	0.0	865.0
	YASA	301.1	94.0	192.2	248.8	302.5	351.0	398.0	0.0	768.5
WASO [*minutes*]	-	64.2	53.8	12.5	25.0	50.0	89.2	135.0	0.0	952.0
	U-Sleep	61.3	52.5	12.5	23.5	46.5	84.0	132.0	0.0	932.0
	YASA	92.8	62.8	29.0	46.5	78.0	124.5	174.0	3.0	980.0
Sleep Cycles [*N*]	-	2.6	1.4	1.0	1.5	2.5	3.5	4.0	0.0	11.0
	U-Sleep	2.8	1.4	1.0	2.0	2.5	3.5	4.5	0.0	10.5
	YASA	2.6	1.2	1.0	1.5	2.5	3.5	4.0	0.0	9.5
Sleep-stage Transitions [*N/hour*]	-	21.2	7.3	12.6	16.1	20.5	25.3	30.8	1.2	64.4
	U-Sleep	14.6	4.9	9.3	11.3	13.9	17.2	20.8	0.9	62.3
	YASA	17.1	5.4	11.2	13.5	16.4	19.9	23.9	0.2	69.8
Awakenings [*N/hour*]	-	3.5	2.2	1.4	2.1	3.1	4.4	6.0	0.0	27.9
	U-Sleep	3.3	1.8	1.5	2.2	3.0	4.2	5.5	0.0	20.6
	YASA	4.9	2.3	2.6	3.3	4.5	6.0	7.9	0.2	28.6
Sleep Efficiency [%]	-	80.1	14.9	60.4	73.4	83.7	91.2	94.9	0.0	100.0
	U-Sleep	81.5	14.4	62.4	75.4	85.3	91.8	95.2	0.0	100.0
	YASA	71.3	17.2	48.4	62.5	74.9	84.5	89.4	0.0	98.1
W [%]	-	19.9	14.9	5.1	8.8	16.3	26.6	39.6	0.0	100.0
	U-Sleep	18.5	14.4	4.8	8.2	14.7	24.6	37.6	0.0	100.0
	YASA	28.7	17.2	10.6	15.5	25.1	37.5	51.6	1.9	100.0
N1 [%]	-	15.9	10.3	5.8	8.7	13.4	20.3	29.1	0.0	85.5
	U-Sleep	10.6	7.8	3.5	5.4	8.5	13.4	20.0	0.0	69.2
	YASA	5.1	3.8	1.2	2.6	4.4	6.8	9.5	0.0	39.2
N2 [%]	-	35.5	12.3	19.2	27.7	36.3	43.9	50.1	0.0	87.4
	U-Sleep	44.5	12.1	29.1	37.8	45.5	52.1	58.5	0.0	96.9
	YASA	39.9	10.1	27.2	34.6	41.0	46.4	51.4	0.0	73.5
N3 [%]	-	16.2	10.3	2.7	9.0	15.4	22.3	29.3	0.0	75.3
	U-Sleep	13.1	8.6	1.4	6.7	12.8	18.7	24.0	0.0	58.6
	YASA	13.7	8.4	1.6	7.3	13.7	19.6	24.7	0.0	48.6
REM [%]	-	12.6	6.9	3.3	7.8	12.5	17.2	21.4	0.0	48.5
	U-Sleep	13.3	7.2	3.4	8.3	13.4	18.1	22.3	0.0	54.9
	YASA	12.6	6.5	3.5	8.0	12.7	17.3	21.0	0.0	44.1

Notes: For each metric, the mean, standard deviation (SD), quantiles (Q10, Q25, Q50, Q75, Q90), minimum (Min), and maximum (Max) are reported. Paired Wilcoxon signed-rank tests compared model predictions against physician-based reference values, with significant results highlighted by model name according to p-value thresholds: 0.05 , 0.01 , and 0.001 .

A.2 Statistical characteristics of raw errors in algorithm-derived PSG markers

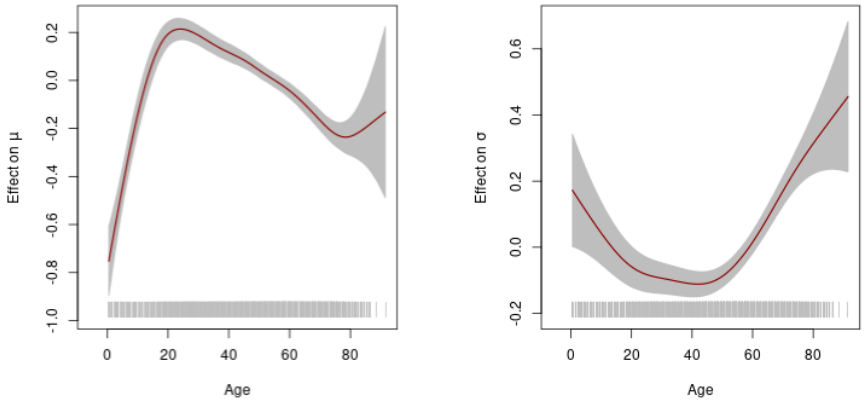
Table A.2: Summary of prediction errors in sleep metrics from U-Sleep and YASA compared to physician scoring.

Metric-Error	Algorithm	Mean	SD	Q10	Q25	Q50	Q75	Q90	Min	Max
Sleep Latency	U-Sleep	-2.5	14.7	-10.0	-2.0	0.0	0.5	3.0	-271.0	160.5
[*minutes*]	YASA	8.5	19.2	0.0	1.0	3.5	10.0	22.5	-266.5	262.5
REM Latency	U-Sleep	-0.3	69.1	-45.6	-2.5	0.0	3.5	44.5	-410.5	608.0
[*minutes*]	YASA	-18.4	79.0	-112.5	-28.0	-4.5	0.0	50.8	-394.5	411.5
Total Sleep Time	U-Sleep	5.2	24.6	-12.0	-3.5	2.0	10.5	25.5	-353.5	263.0
[*minutes*]	YASA	-37.3	42.3	-81.5	-47.0	-26.0	-14.0	-6.0	-532.5	164.5
WASO	U-Sleep	-2.8	24.1	-21.5	-8.5	-1.5	4.0	14.5	-259.0	317.5
[*minutes*]	YASA	28.8	38.5	1.0	9.0	20.0	38.8	68.0	-242.0	426.5
Sleep Cycles	U-Sleep	0.2	0.7	0.0	0.0	0.0	0.0	1.0	-8.0	5.0
[*N*]	YASA	-0.0	0.9	-1.0	0.0	0.0	0.0	1.0	-10.5	3.5
Sleep-stage Transitions	U-Sleep	-6.5	6.0	-13.7	-9.9	-6.2	-3.0	-0.1	-39.8	36.0
[*N/hour*]	YASA	-4.0	6.4	-11.5	-7.8	-3.9	-0.3	3.4	-38.2	37.5
Awakenings	U-Sleep	-0.1	1.5	-1.5	-0.7	-0.1	0.5	1.3	-17.1	9.5
[*N/hour*]	YASA	1.4	2.2	-0.8	0.2	1.2	2.5	4.2	-12.7	15.6
Sleep Efficiency	U-Sleep	1.3	5.7	-2.7	-0.8	0.6	2.6	6.2	-66.3	63.4
[%]	YASA	-8.8	9.8	-19.4	-11.3	-6.2	-3.3	-1.5	-99.9	40.2
W	U-Sleep	-1.3	5.7	-6.2	-2.6	-0.6	0.8	2.7	-63.4	66.3
[%]	YASA	8.8	9.8	1.5	3.3	6.2	11.3	19.4	-40.2	99.9
N1	U-Sleep	-5.4	7.6	-14.4	-8.6	-4.1	-1.0	1.5	-61.5	41.1
[%]	YASA	-10.8	10.0	-23.4	-15.0	-8.6	-4.4	-1.7	-77.4	34.6
N2	U-Sleep	9.0	9.1	0.0	3.2	7.5	13.1	20.0	-30.1	78.3
[%]	YASA	4.5	8.9	-5.0	-0.7	3.7	9.0	15.3	-78.1	51.6
N3	U-Sleep	-3.1	6.1	-10.2	-5.7	-2.2	0.0	2.8	-75.3	30.6
[%]	YASA	-2.5	6.0	-9.7	-4.8	-1.4	0.5	3.2	-75.3	23.0
REM	U-Sleep	0.7	3.1	-1.9	-0.3	0.5	1.9	3.8	-37.6	35.3
[%]	YASA	0.0	4.1	-4.2	-1.5	0.3	2.1	4.1	-39.8	25.0

Notes: The table reports the mean error, standard deviation (SD), quantiles (Q10, Q25, Q50, Q75, Q90), minimum (Min), and maximum (Max). Paired Wilcoxon signed-rank tests compared model predictions against physician-based reference values, testing whether differences were symmetrically distributed around zero. Significant results highlight the corresponding model name (U-Sleep or YASA) according to p-value thresholds: 0.05 , 0.01 , and 0.001 .

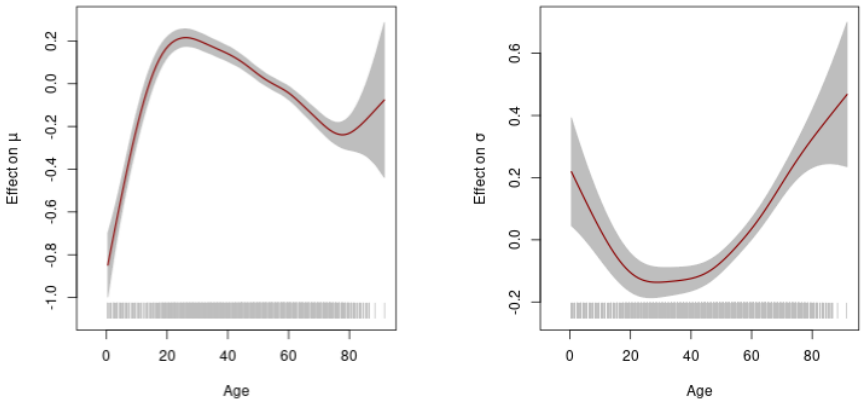
A.3 Partial effects of age on U-Sleep and YASA performance metrics

Figure A.1: Partial effects of age on U-Sleep accuracy.



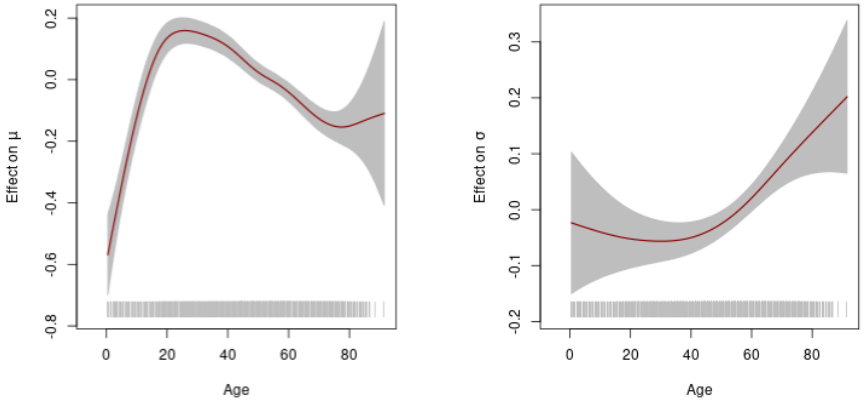
Notes: The left panel shows the estimated expected bias (location parameter) across ages, while the right panel illustrates its variability (scale parameter). Shaded areas represent 95% confidence intervals.

Figure A.2: Partial effects of age on U-Sleep F1-score.



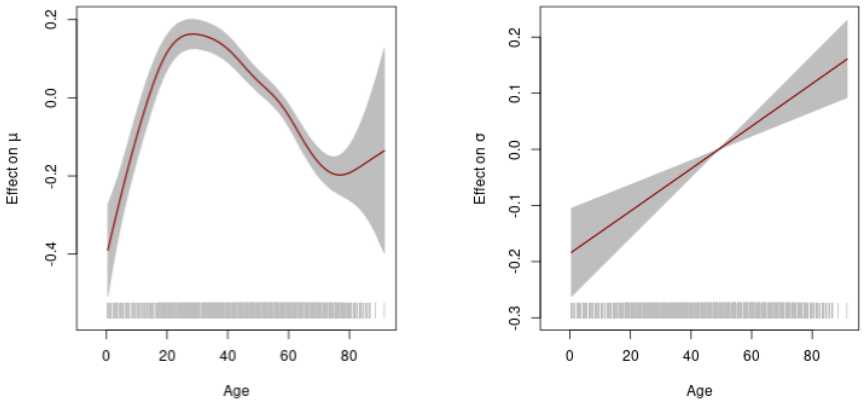
Notes: The left panel shows the estimated expected bias (location parameter) across ages, while the right panel illustrates its variability (scale parameter). Shaded areas represent 95% confidence intervals.

Figure A.3: Partial effects of age on YASA accuracy.



Notes: The left panel shows the estimated expected bias (location parameter) across ages, while the right panel illustrates its variability (scale parameter). Shaded areas represent 95% confidence intervals.

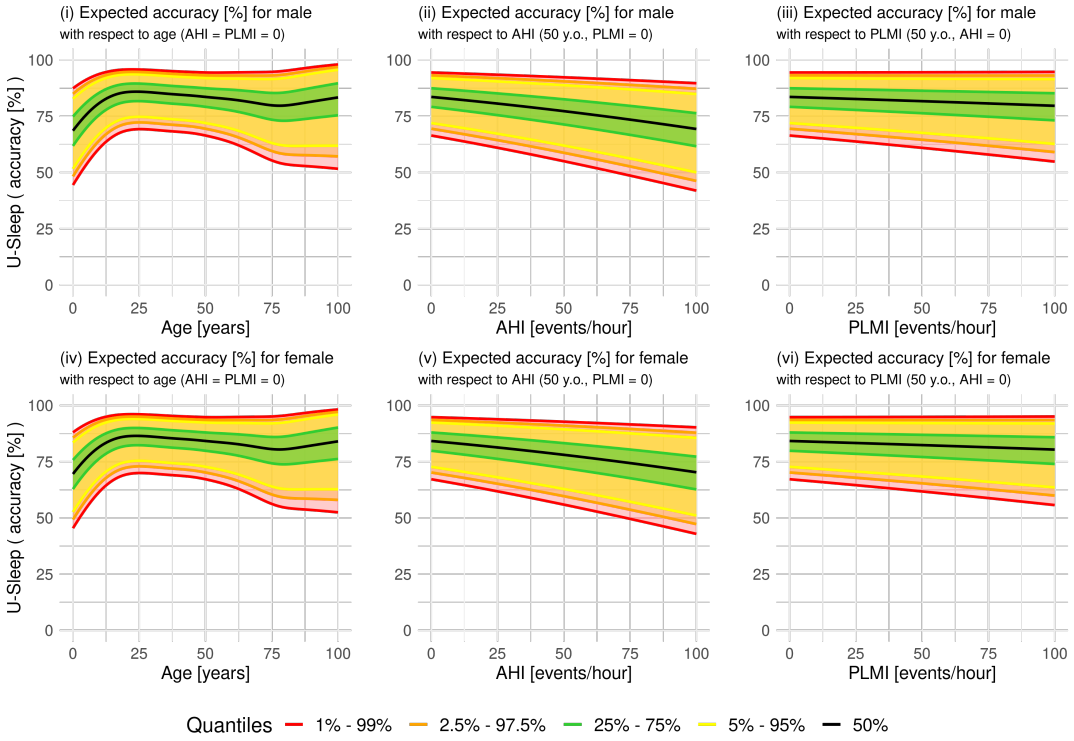
Figure A.4: Partial effects of age on YASA F1-score.



Notes: The left panel shows the estimated expected bias (location parameter) across ages, while the right panel illustrates its variability (scale parameter). Shaded areas represent 95% confidence intervals.

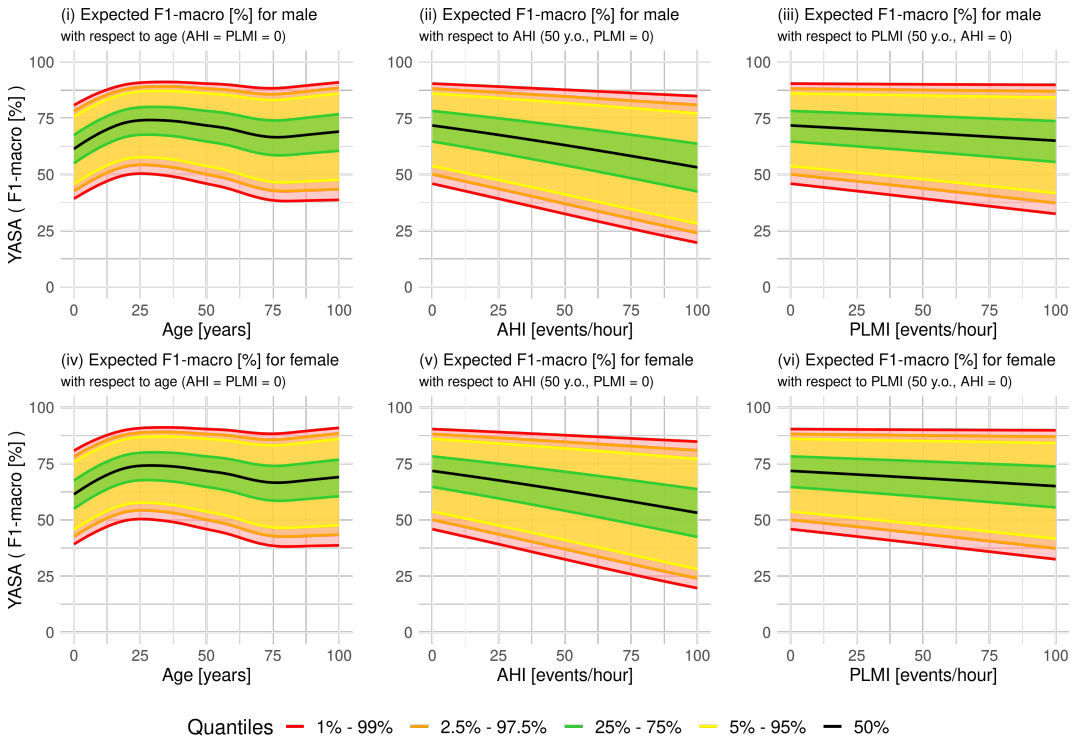
A.4 Performance Plots

Figure A.5: Expected distribution of subject-specific accuracy for U-Sleep across demographic and clinical subgroups.



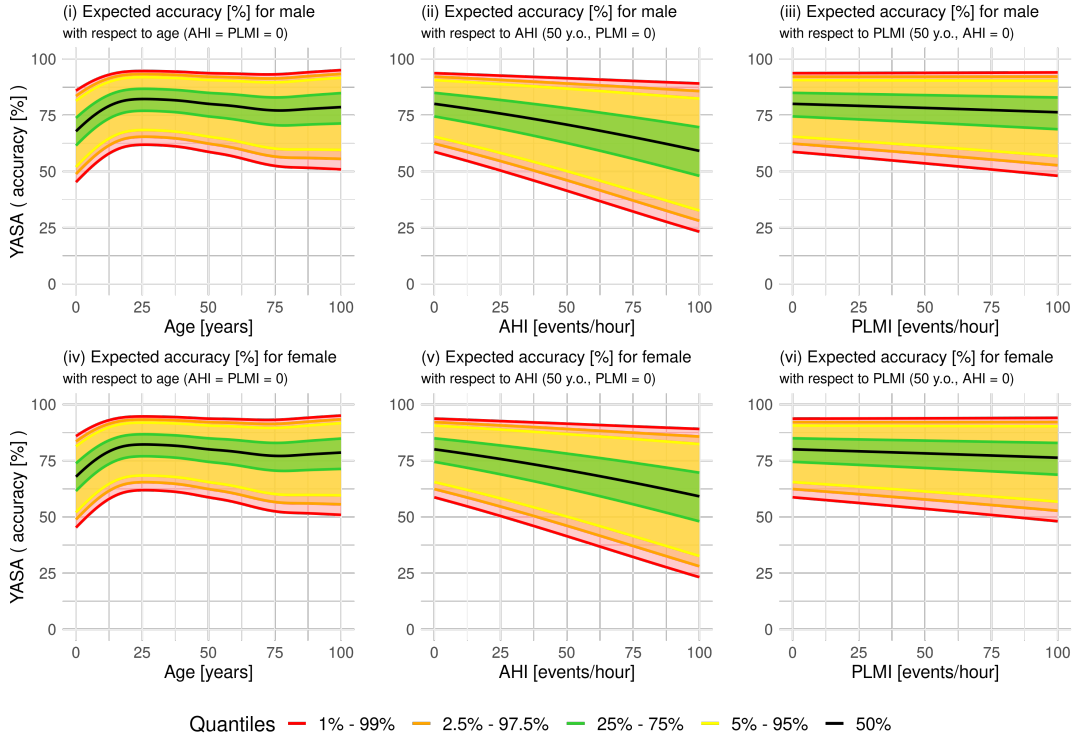
Notes: Expected distribution of the subject-specific accuracy based on the zero-and-ones-inflated Beta performance model for U-Sleep predictions, stratified by gender (top row: males, bottom row: females). The graphs display the estimated distribution as a function of bias-inducing variables (age, AHI, and PLMI) on the horizontal axis. The solid black line represents the median, while the shaded areas correspond to different percentile ranges: 25-75% in green, 5-95% in yellow, 2.5-97.5% in orange, and 1-99% in red, showing the expected performance variability across subjects' characteristics.

Figure A.6: Expected distribution of subject-specific F1-score for YASA across demographic and clinical subgroups.



Notes: Expected distribution of the subject-specific macro F1-score based on the zero-and-ones-inflated Beta performance model for YASA predictions, stratified by gender (top row: males, bottom row: females). The graphs display the estimated distribution as a function of bias-inducing variables (age, AHI, and PLMI) on the horizontal axis. The solid black line represents the median, while the shaded areas correspond to different percentile ranges: 25-75% in green, 5-95% in yellow, 2.5-97.5% in orange, and 1-99% in red, reflecting the expected performance variability across subjects' characteristics.

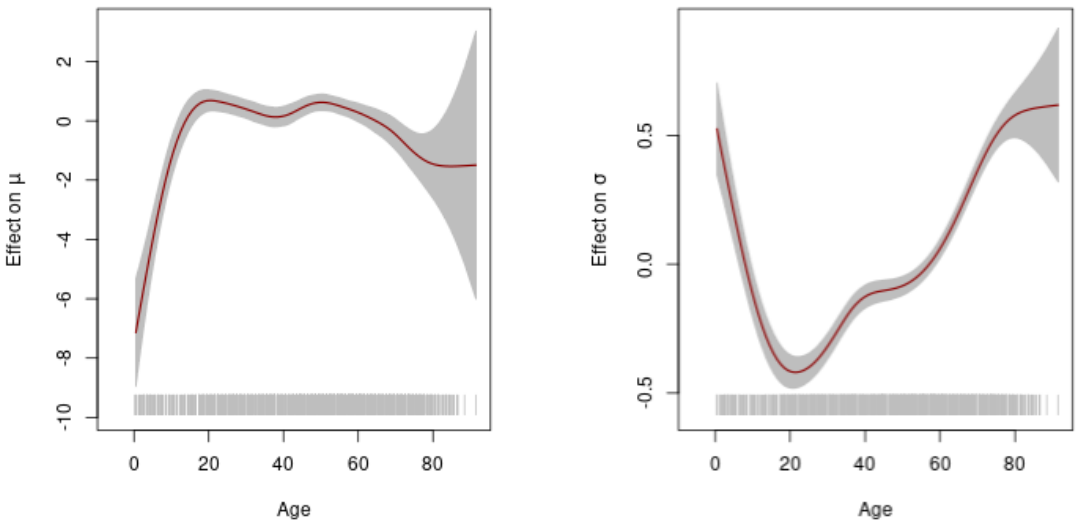
Figure A.7: Expected distribution of subject-specific accuracy for YASA across demographic and clinical subgroups.



Notes: Expected distribution of the subject-specific accuracy based on the zero-and-ones-inflated Beta performance model for YASA predictions, stratified by gender (top row: males, bottom row: females). The graphs display the estimated distribution as a function of bias-inducing variables (age, AHI, and PLMI) on the horizontal axis. The solid black line represents the median, while the shaded areas correspond to different percentile ranges: 25-75% in green, 5-95% in yellow, 2.5-97.5% in orange, and 1-99% in red, showing the expected performance variability across subjects' characteristics.

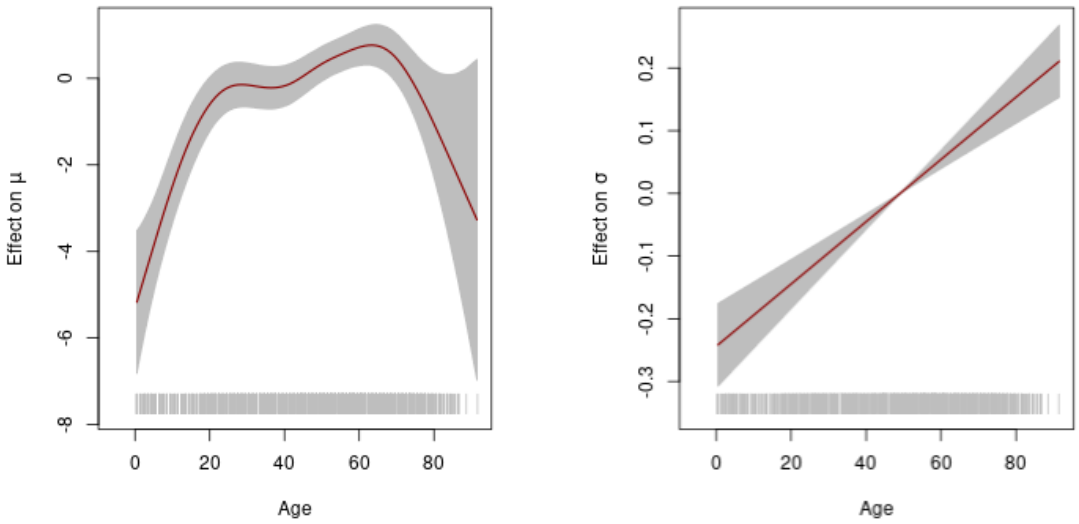
A.5 Partial effects of age on bias in U-Sleep and YASA derived percentage of wakefulness

Figure A.8: Partial effects of age on bias in U-Sleep-derived wakefulness percentage (W%).



Notes: The left panel shows the estimated expected bias (location parameter) across ages, while the right panel illustrates its variability (scale parameter). Shaded areas represent 95% confidence intervals.

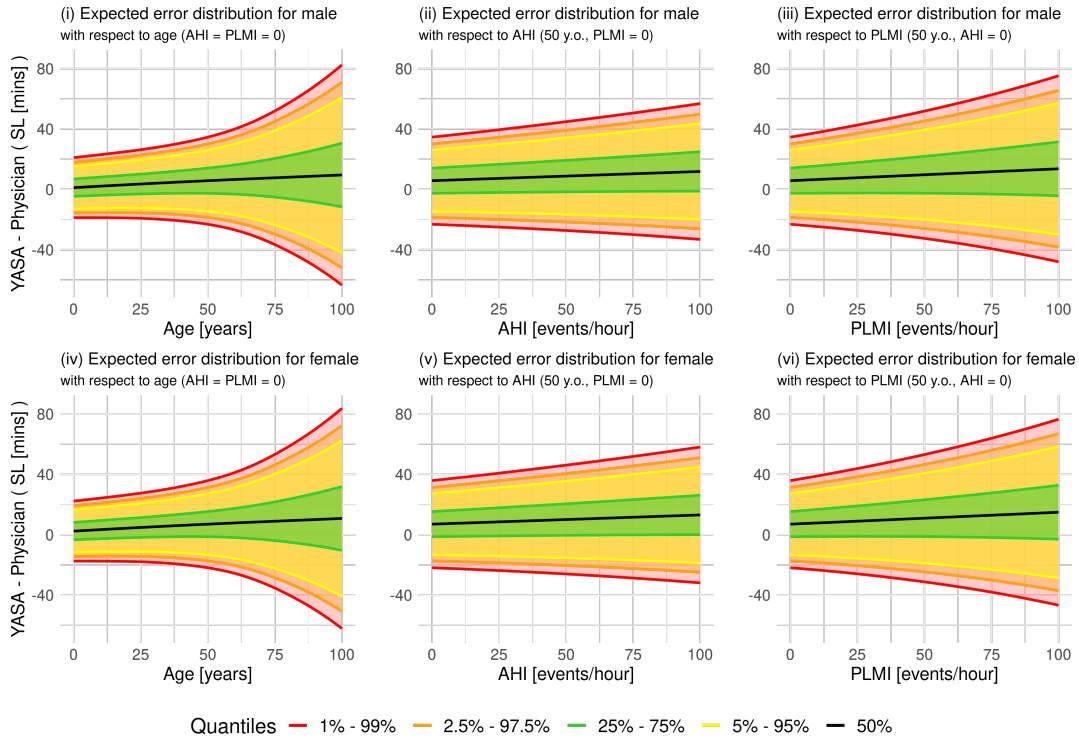
Figure A.9: Partial effects of age on bias in YASA-derived wakefulness percentage (W%).



Notes: The left panel shows the estimated expected bias (location parameter) across ages, while the right panel illustrates its variability (scale parameter). Shaded areas represent 95% confidence intervals.

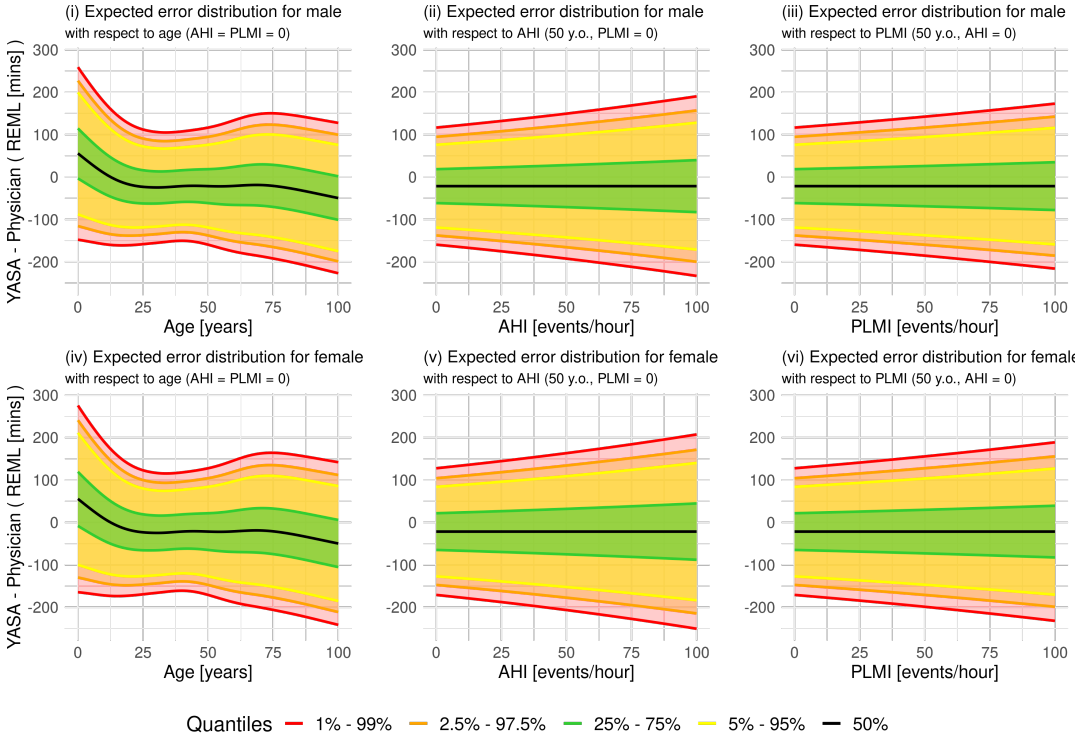
A.6 Bias in clinical PSG markers based on YASA predictions

Figure A.10: Expected distribution of the bias in the sleep latency (SL, minutes) for YASA predictions.



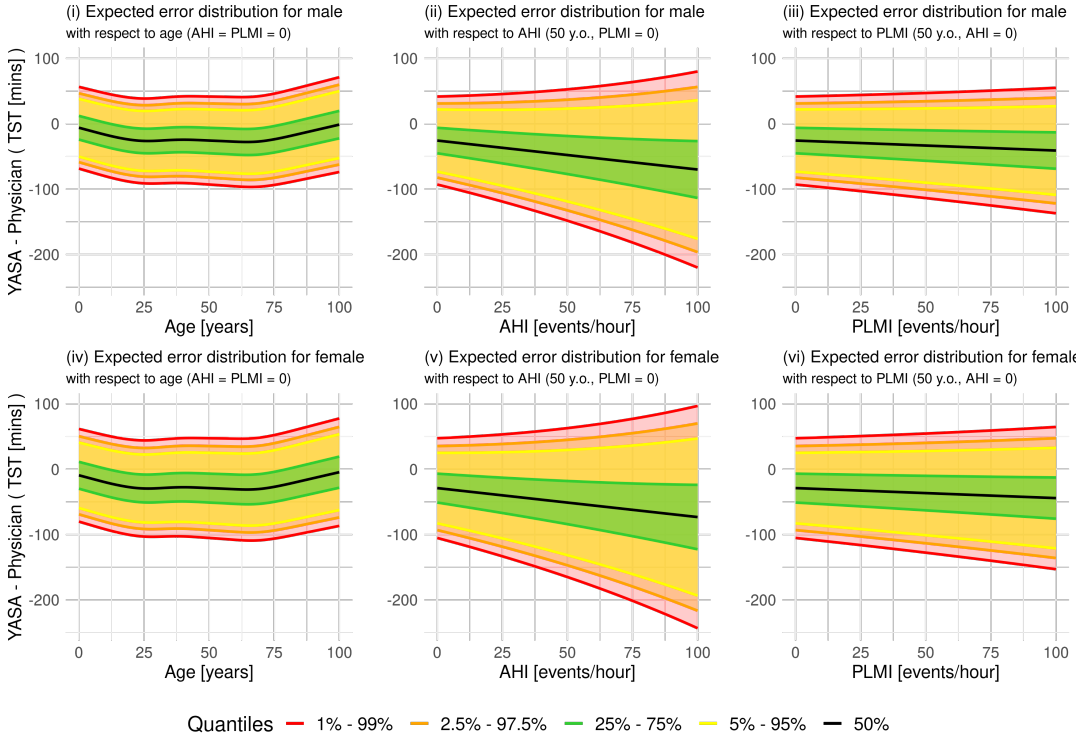
Notes: Expected distribution of the bias in the sleep latency (SL, minutes) based on the generalized normal distribution for YASA predictions, stratified by gender (top row: males, bottom row: females). The graphs display the estimated distribution as a function of bias-inducing variables (age, AHI, and PLMI) on the horizontal axis. The solid black line represents the median, while the shaded areas correspond to different percentile ranges: 25-75% in green, 5-95% in yellow, 2.5-97.5% in orange, and 1-99% in red, illustrating the expected performance variability across subjects' characteristics.

Figure A.11: Expected distribution of the bias in the REM latency (REML, minutes) for YASA predictions.



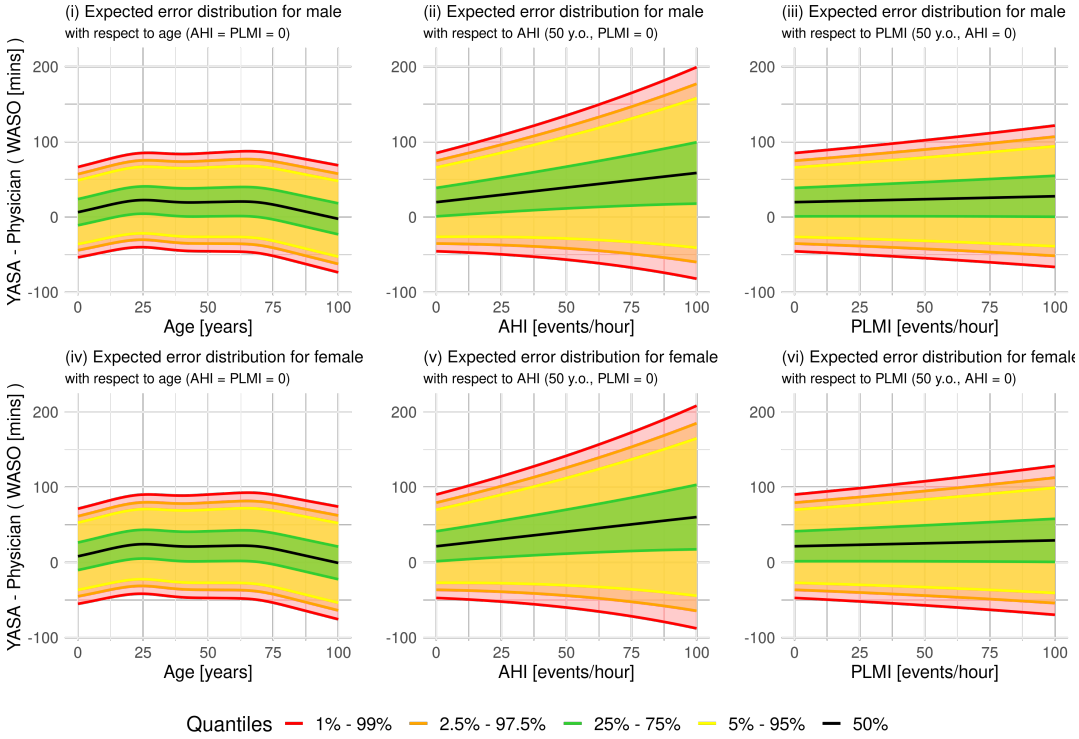
Notes: Expected distribution of the bias in the REM latency (REML, minutes) based on the generalized normal distribution for YASA predictions, stratified by gender (top row: males, bottom row: females). The graphs display the estimated distribution as a function of bias-inducing variables (age, AHI, and PLMI) on the horizontal axis. The solid black line represents the median, while the shaded areas correspond to different percentile ranges: 25-75% in green, 5-95% in yellow, 2.5-97.5% in orange, and 1-99% in red, illustrating the expected performance variability across subjects' characteristics.

Figure A.12: Expected distribution of the bias in the total sleep time (TST, minutes) for YASA predictions.



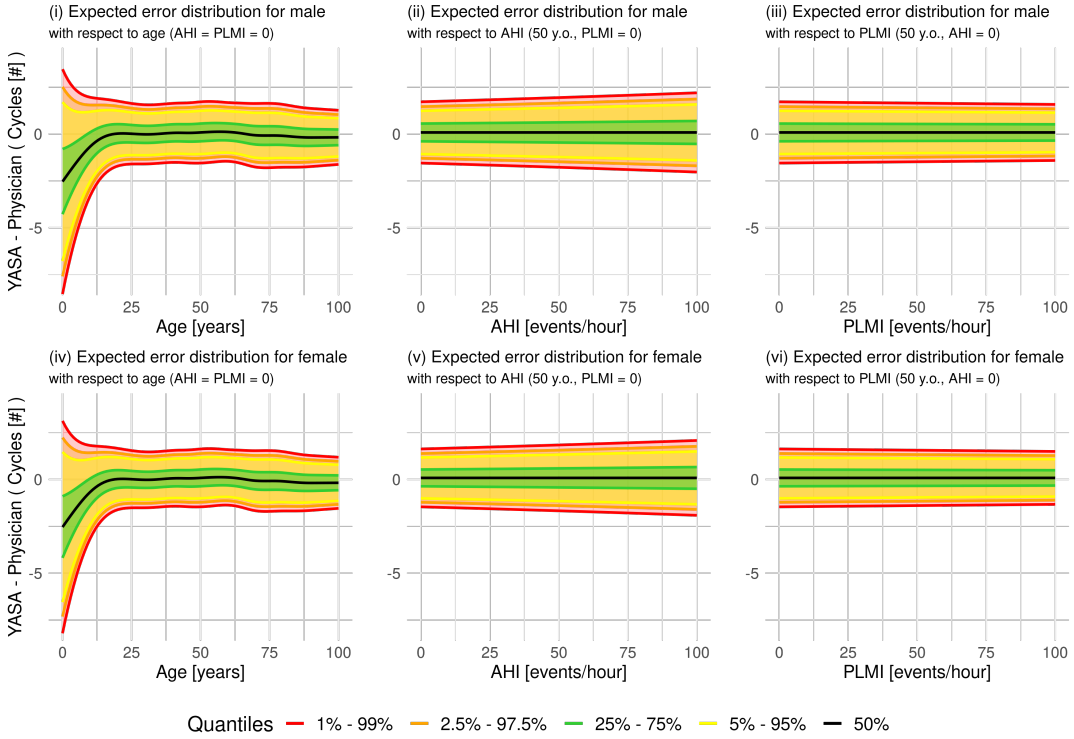
Notes: Expected distribution of the bias in the total sleep time (TST, minutes) based on the generalized normal distribution for YASA predictions, stratified by gender (top row: males, bottom row: females). The graphs display the estimated distribution as a function of bias-inducing variables (age, AHI, and PLMI) on the horizontal axis. The solid black line represents the median, while the shaded areas correspond to different percentile ranges: 25-75% in green, 5-95% in yellow, 2.5-97.5% in orange, and 1-99% in red, illustrating the expected performance variability across subjects' characteristics.

Figure A.13: Expected distribution of the bias in the wake after sleep onset (WASO, minutes) for YASA predictions.



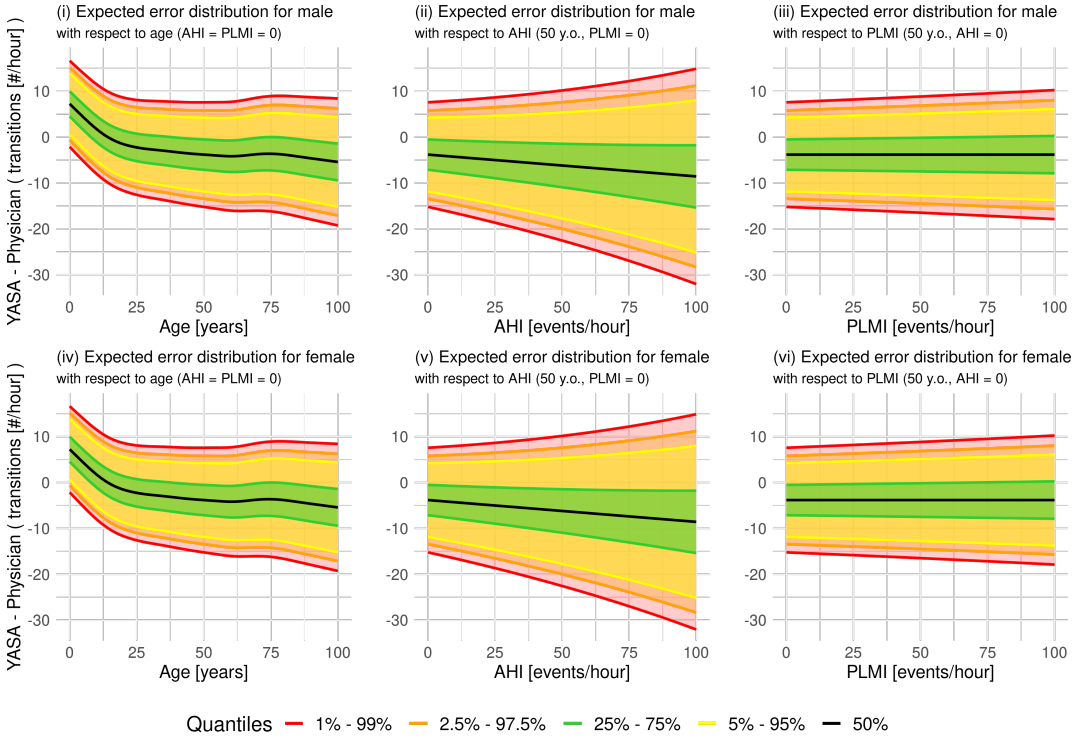
Notes: Expected distribution of the bias in the wake after sleep onset (WASO, minutes) based on the generalized normal distribution for YASA predictions, stratified by gender (top row: males, bottom row: females). The graphs display the estimated distribution as a function of bias-inducing variables (age, AHI, and PLMI) on the horizontal axis. The solid black line represents the median, while the shaded areas correspond to different percentile ranges: 25-75% in green, 5-95% in yellow, 2.5-97.5% in orange, and 1-99% in red, illustrating the expected performance variability across subjects' characteristics.

Figure A.14: Expected distribution of the bias in the number (#) of sleep cycles for YASA predictions.



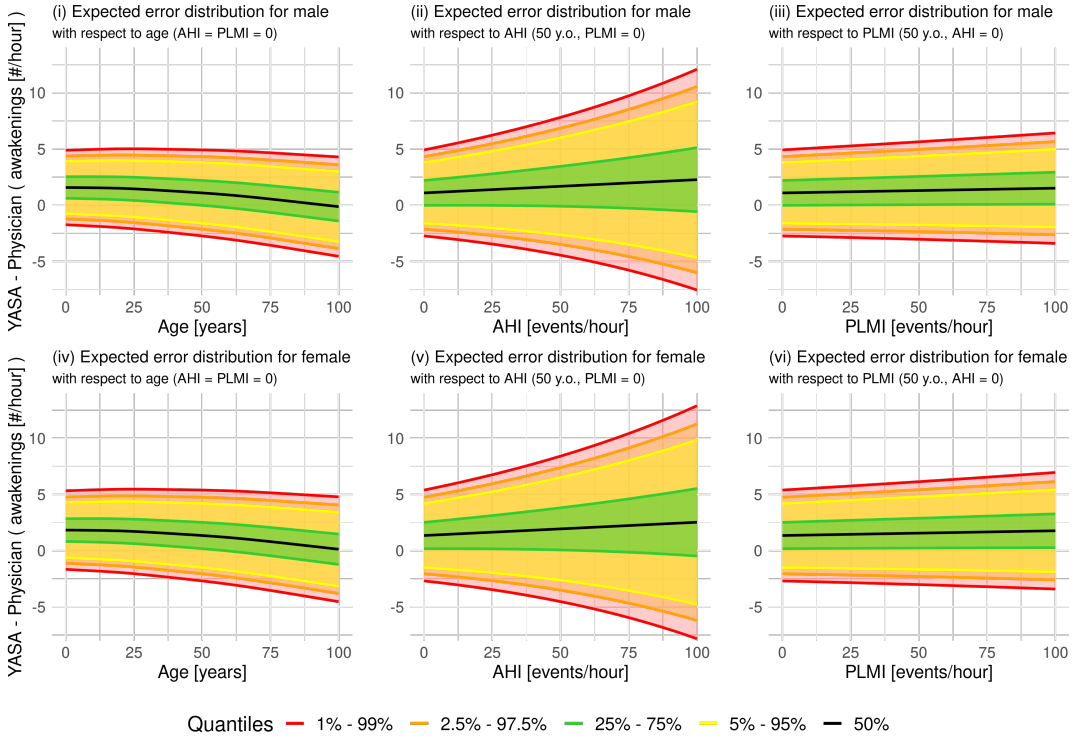
Notes: Expected distribution of the bias in the number (#) of sleep cycles based on the generalized normal distribution for YASA predictions, stratified by gender (top row: males, bottom row: females). The graphs display the estimated distribution as a function of bias-inducing variables (age, AHI, and PLMI) on the horizontal axis. The solid black line represents the median, while the shaded areas correspond to different percentile ranges: 25-75% in green, 5-95% in yellow, 2.5-97.5% in orange, and 1-99% in red, illustrating the expected performance variability across subjects' characteristics.

Figure A.15: Expected distribution of the bias in the hourly rate (# / hour) of sleep stage transitions for YASA predictions.



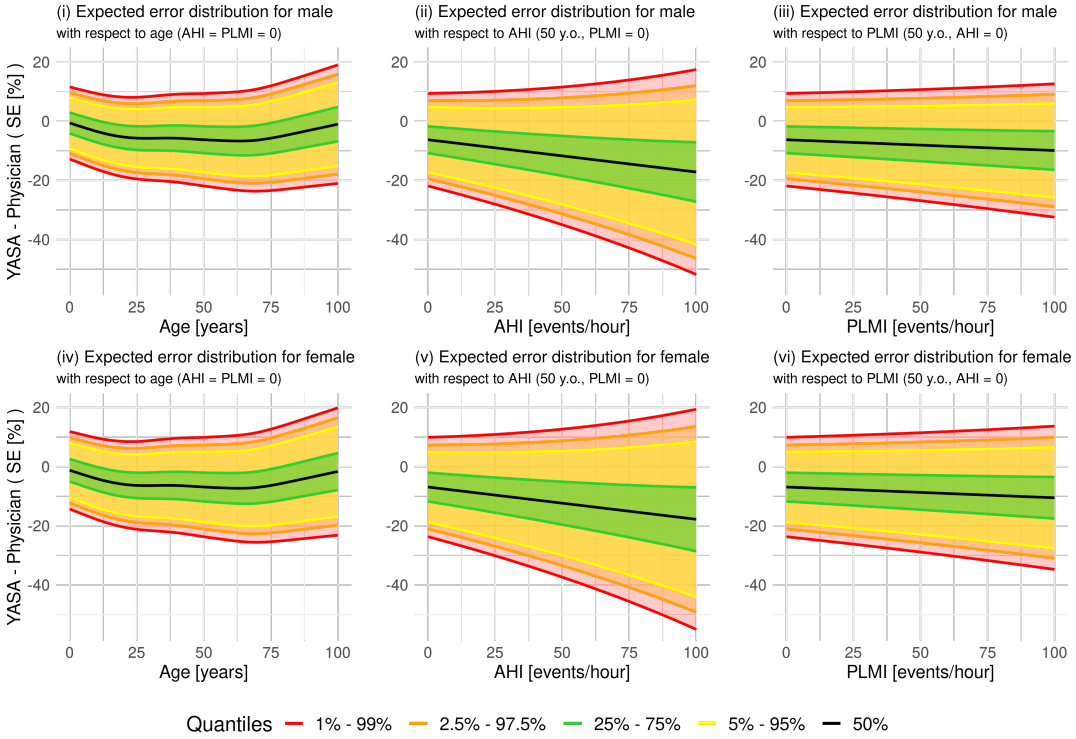
Notes: Expected distribution of the bias in the hourly rate (# / hour) of sleep stage transitions based on the generalized normal distribution for YASA predictions, stratified by gender (top row: males, bottom row: females). The graphs display the estimated distribution as a function of bias-inducing variables (age, AHI, and PLMI) on the horizontal axis. The solid black line represents the median, while the shaded areas correspond to different percentile ranges: 25-75% in green, 5-95% in yellow, 2.5-97.5% in orange, and 1-99% in red, illustrating the expected performance variability across subjects' characteristics.

Figure A.16: Expected distribution of the bias in the hourly rate (# / hour) of awakenings for YASA predictions.



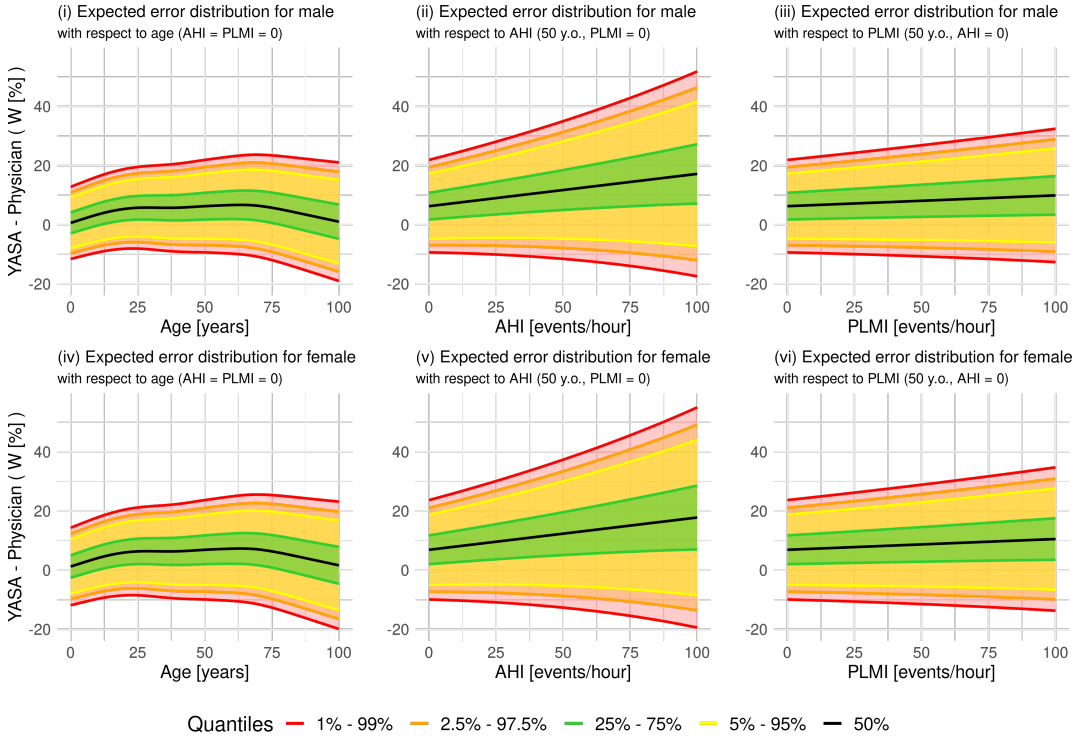
Notes: Expected distribution of the bias in the hourly rate (# / hour) of awakenings based on the generalized normal distribution for YASA predictions, stratified by gender (top row: males, bottom row: females). The graphs display the estimated distribution as a function of bias-inducing variables (age, AHI, and PLMI) on the horizontal axis. The solid black line represents the median, while the shaded areas correspond to different percentile ranges: 25-75% in green, 5-95% in yellow, 2.5-97.5% in orange, and 1-99% in red, illustrating the expected performance variability across subjects' characteristics.

Figure A.17: Expected distribution of the bias in the sleep efficiency percentage (SE, %) for YASA predictions.



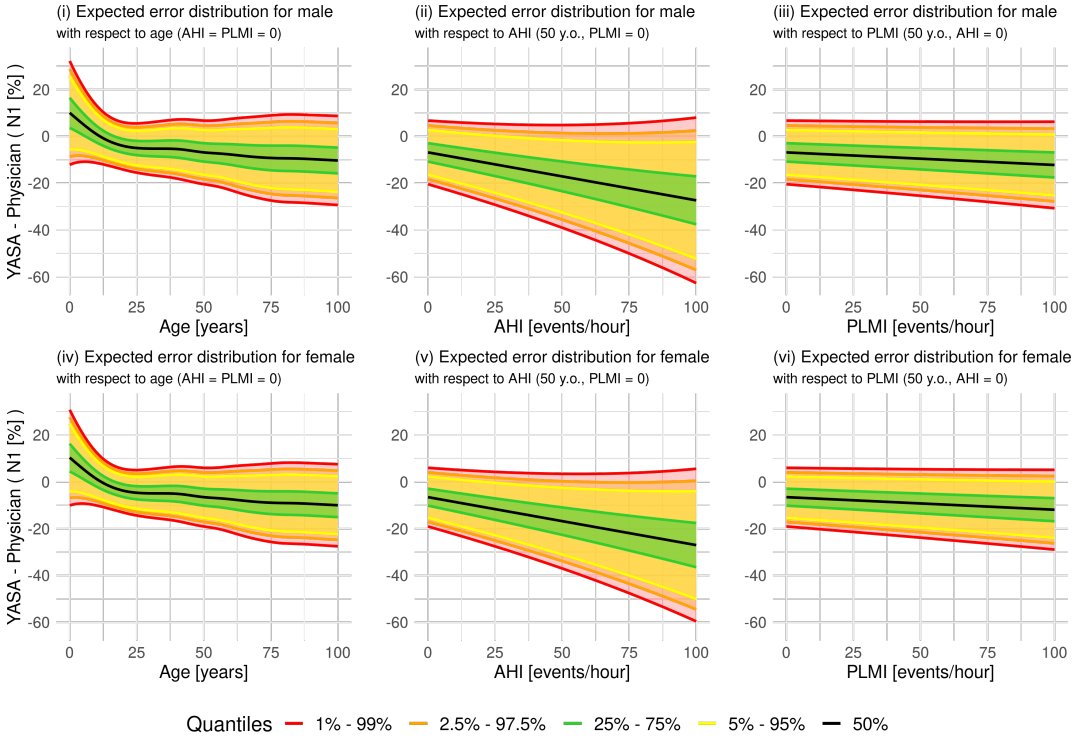
Notes: Expected distribution of the bias in the sleep efficiency percentage (SE, %) based on the generalized normal distribution for YASA predictions, stratified by gender (top row: males, bottom row: females). The graphs display the estimated distribution as a function of bias-inducing variables (age, AHI, and PLMI) on the horizontal axis. The solid black line represents the median, while the shaded areas correspond to different percentile ranges: 25-75% in green, 5-95% in yellow, 2.5-97.5% in orange, and 1-99% in red, illustrating the expected performance variability across subjects' characteristics.

Figure A.18: Expected distribution of the bias in the wakefulness percentage after sleep onset (W, %) for YASA predictions.



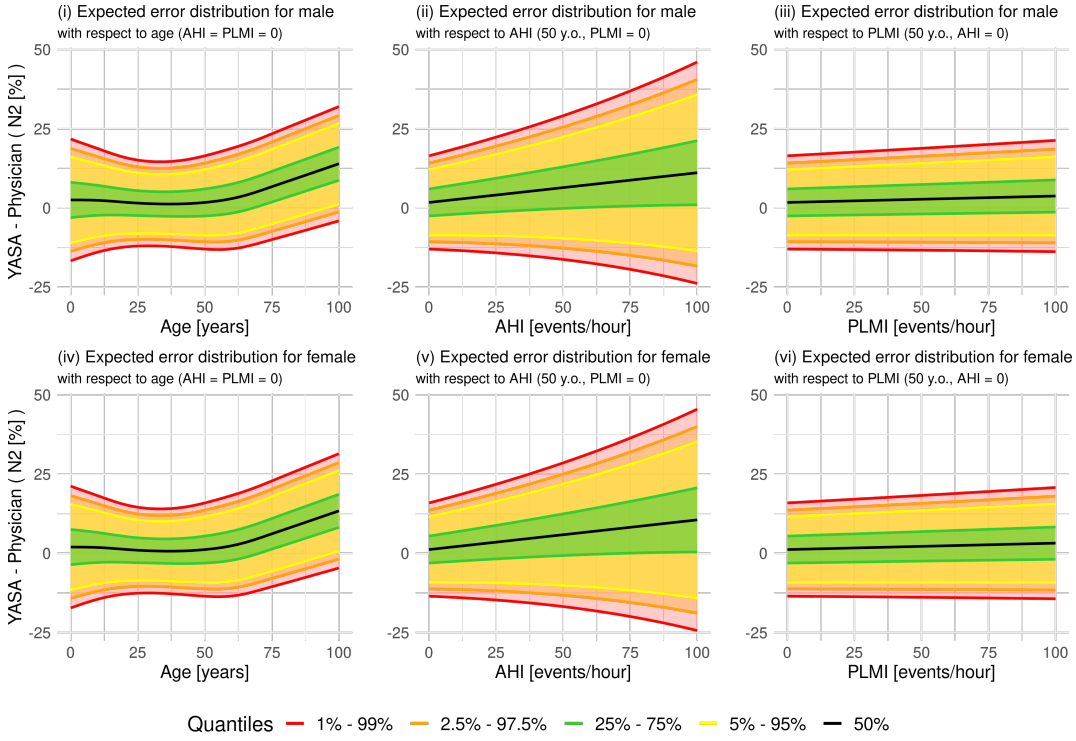
Notes: Expected distribution of the bias in the wakefulness percentage after sleep onset (W, %) based on the generalized normal distribution for YASA predictions, stratified by gender (top row: males, bottom row: females). The graphs display the estimated distribution as a function of bias-inducing variables (age, AHI, and PLMI) on the horizontal axis. The solid black line represents the median, while the shaded areas correspond to different percentile ranges: 25-75% in green, 5-95% in yellow, 2.5-97.5% in orange, and 1-99% in red, illustrating the expected performance variability across subjects' characteristics.

Figure A.19: Expected distribution of the bias in the N1 sleep percentage after sleep onset (N1, %) for YASA predictions.



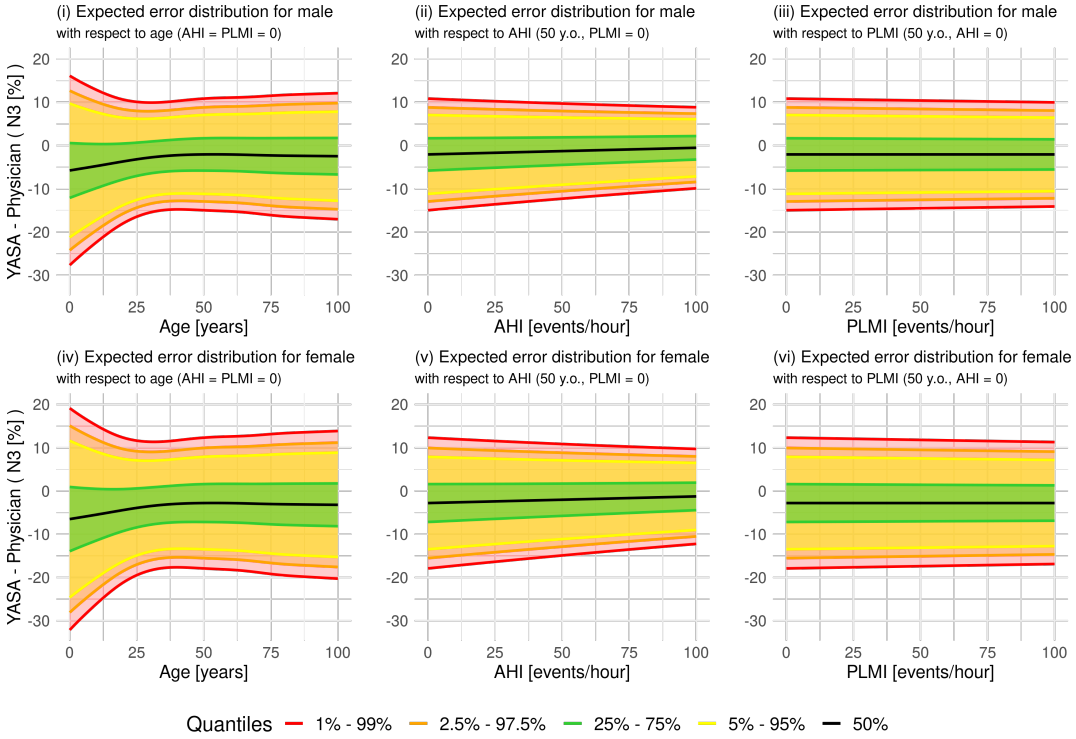
Notes: Expected distribution of the bias in the N1 sleep percentage after sleep onset (N1, %) based on the generalized normal distribution for YASA predictions, stratified by gender (top row: males, bottom row: females). The graphs display the estimated distribution as a function of bias-inducing variables (age, AHI, and PLMI) on the horizontal axis. The solid black line represents the median, while the shaded areas correspond to different percentile ranges: 25-75% in green, 5-95% in yellow, 2.5-97.5% in orange, and 1-99% in red, illustrating the expected performance variability across subjects' characteristics.

Figure A.20: Expected distribution of the bias in the N2 sleep percentage after sleep onset (N2, %) for YASA predictions.



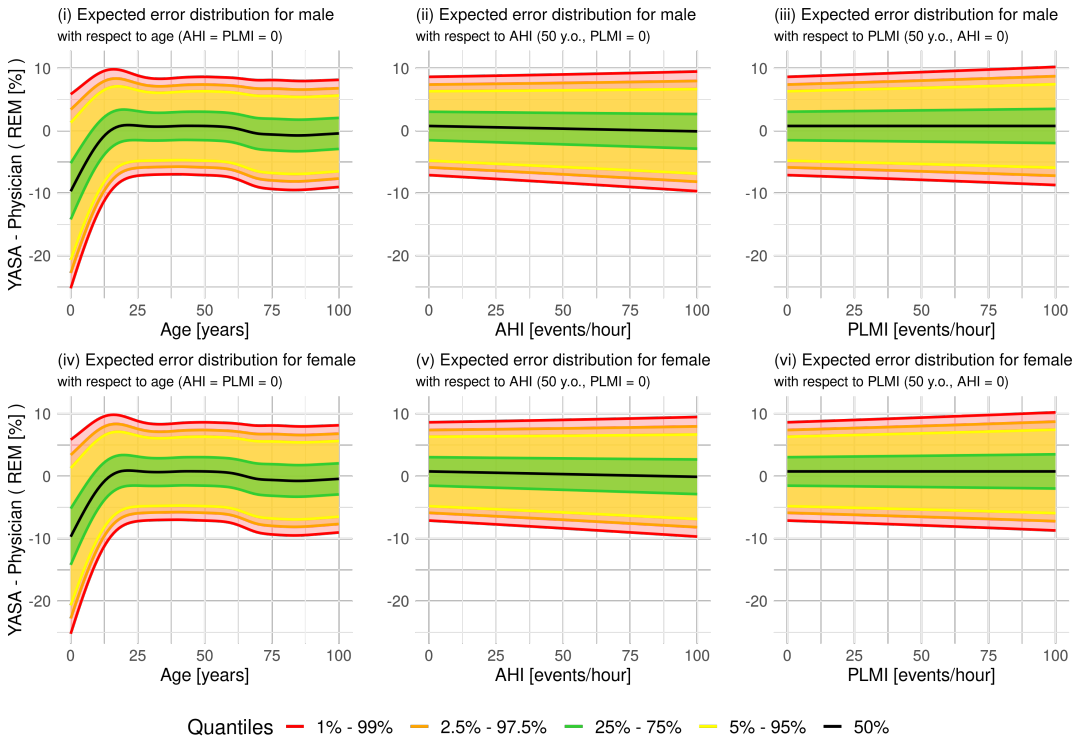
Notes: Expected distribution of the bias in the N2 sleep percentage after sleep onset (N2, %) based on the generalized normal distribution for YASA predictions, stratified by gender (top row: males, bottom row: females). The graphs display the estimated distribution as a function of bias-inducing variables (age, AHI, and PLMI) on the horizontal axis. The solid black line represents the median, while the shaded areas correspond to different percentile ranges: 25-75% in green, 5-95% in yellow, 2.5-97.5% in orange, and 1-99% in red, illustrating the expected performance variability across subjects' characteristics.

Figure A.21: Expected distribution of the bias in the N3 sleep percentage after sleep onset (N3, %) for YASA predictions.



Notes: Expected distribution of the bias in the N3 sleep percentage after sleep onset (N3, %) based on the generalized normal distribution for YASA predictions, stratified by gender (top row: males, bottom row: females). The graphs display the estimated distribution as a function of bias-inducing variables (age, AHI, and PLMI) on the horizontal axis. The solid black line represents the median, while the shaded areas correspond to different percentile ranges: 25-75% in green, 5-95% in yellow, 2.5-97.5% in orange, and 1-99% in red, illustrating the expected performance variability across subjects' characteristics.

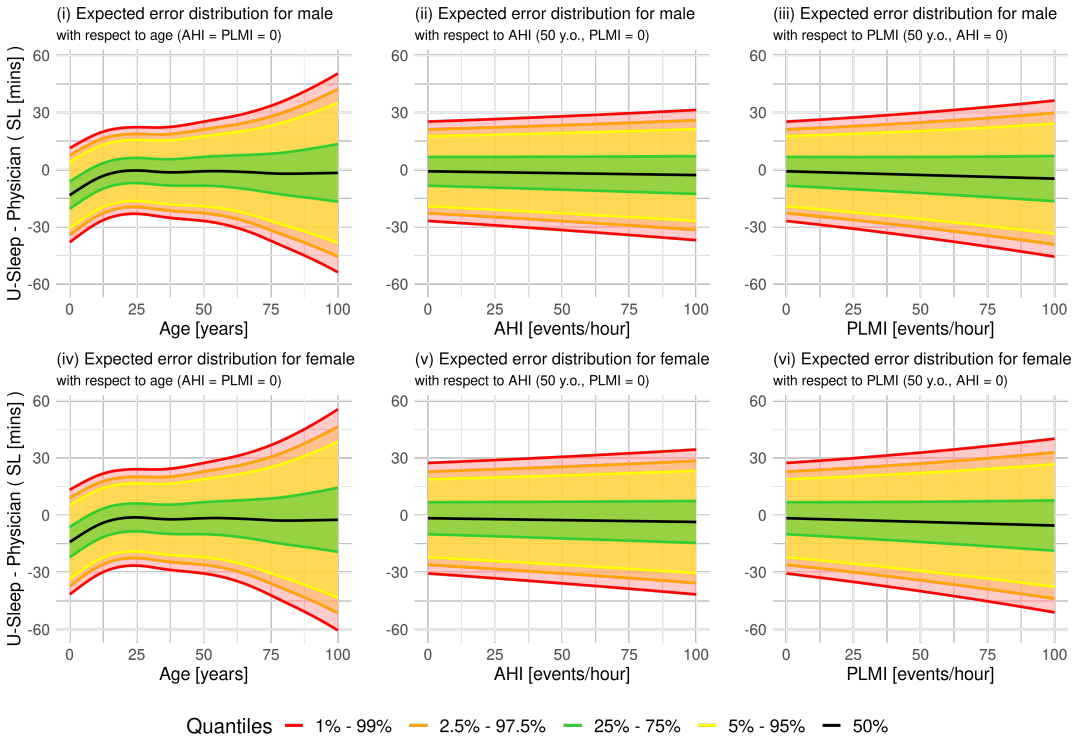
Figure A.22: Expected distribution of the bias in the REM sleep percentage after sleep onset (REM, %) for YASA predictions.



Notes: Expected distribution of the bias in the REM sleep percentage after sleep onset (REM, %) based on the generalized normal distribution for YASA predictions, stratified by gender (top row: males, bottom row: females). The graphs display the estimated distribution as a function of bias-inducing variables (age, AHI, and PLMI) on the horizontal axis. The solid black line represents the median, while the shaded areas correspond to different percentile ranges: 25-75% in green, 5-95% in yellow, 2.5-97.5% in orange, and 1-99% in red, illustrating the expected performance variability across subjects' characteristics.

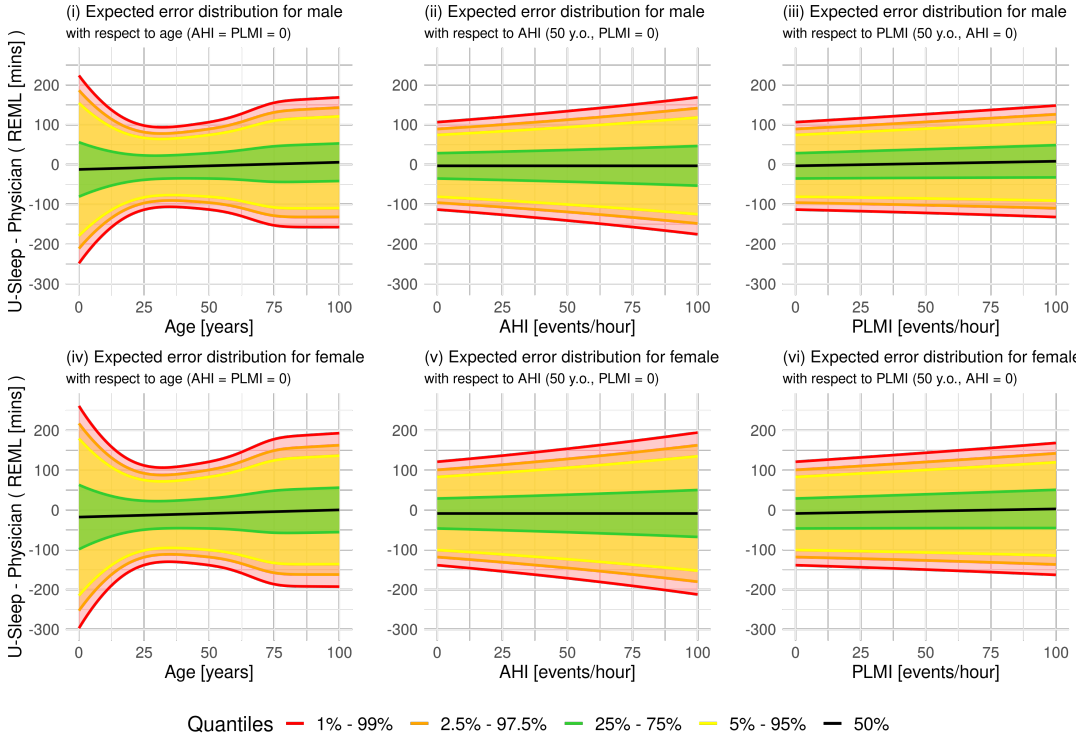
A.7 Bias in clinical PSG markers based on U-Sleep predictions

Figure A.23: Expected distribution of the bias in the sleep latency (SL, minutes) for U-Sleep predictions.



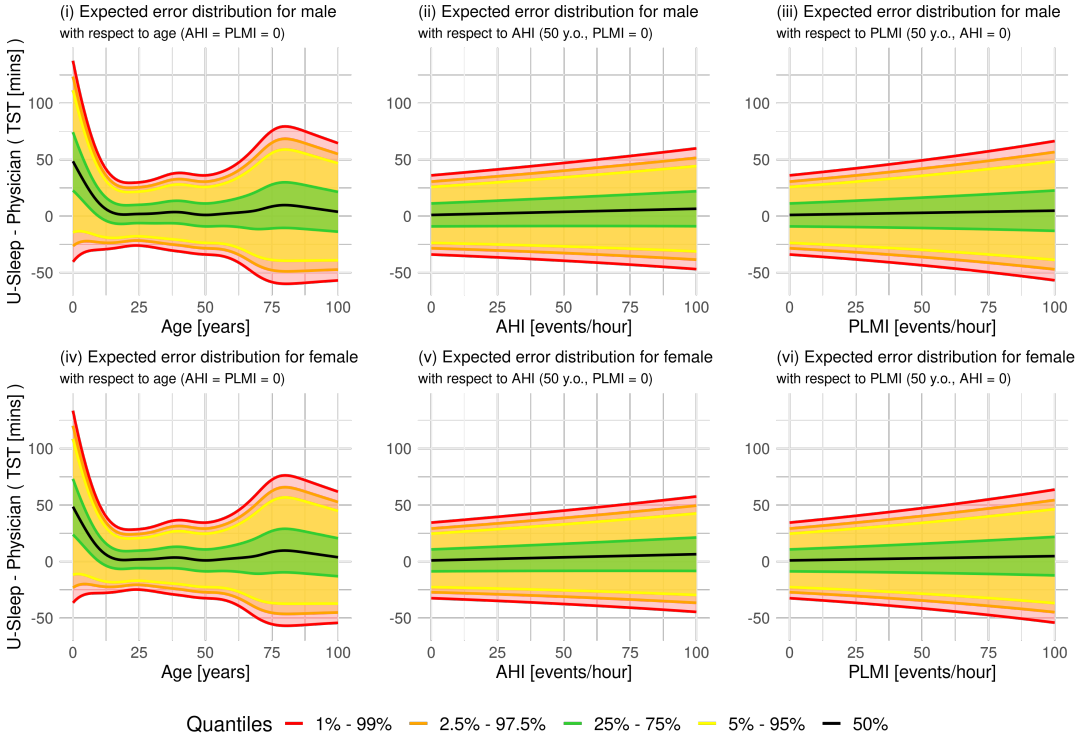
Notes: Expected distribution of the bias in the sleep latency (SL, minutes) based on the generalized normal distribution for U-Sleep predictions, stratified by gender (top row: males, bottom row: females). The graphs display the estimated distribution as a function of bias-inducing variables (age, AHI, and PLMI) on the horizontal axis. The solid black line represents the median, while the shaded areas correspond to different percentile ranges: 25-75% in green, 5-95% in yellow, 2.5-97.5% in orange, and 1-99% in red, illustrating the expected performance variability across subjects' characteristics.

Figure A.24: Expected distribution of the bias in the REM latency (REML, minutes) for U-Sleep predictions.



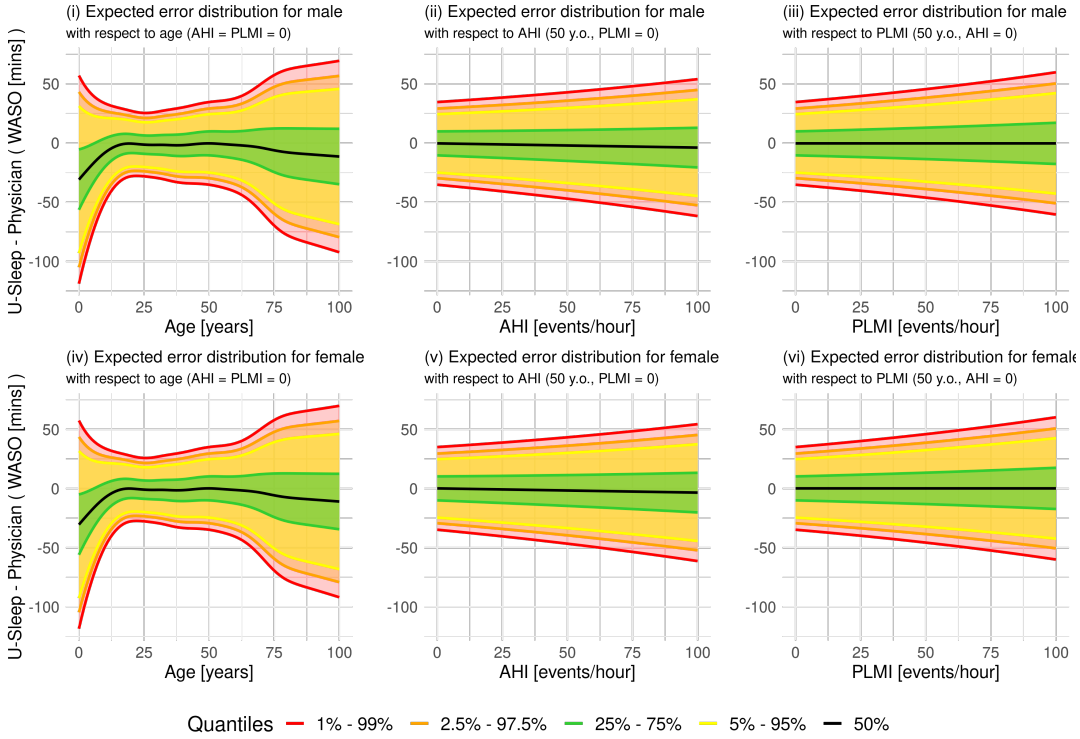
Notes: Expected distribution of the bias in the REM latency (REML, minutes) based on the generalized normal distribution for U-Sleep predictions, stratified by gender (top row: males, bottom row: females). The graphs display the estimated distribution as a function of bias-inducing variables (age, AHI, and PLMI) on the horizontal axis. The solid black line represents the median, while the shaded areas correspond to different percentile ranges: 25-75% in green, 5-95% in yellow, 2.5-97.5% in orange, and 1-99% in red, illustrating the expected performance variability across subjects' characteristics.

Figure A.25: Expected distribution of the bias in the total sleep time (TST, minutes) for U-Sleep predictions.



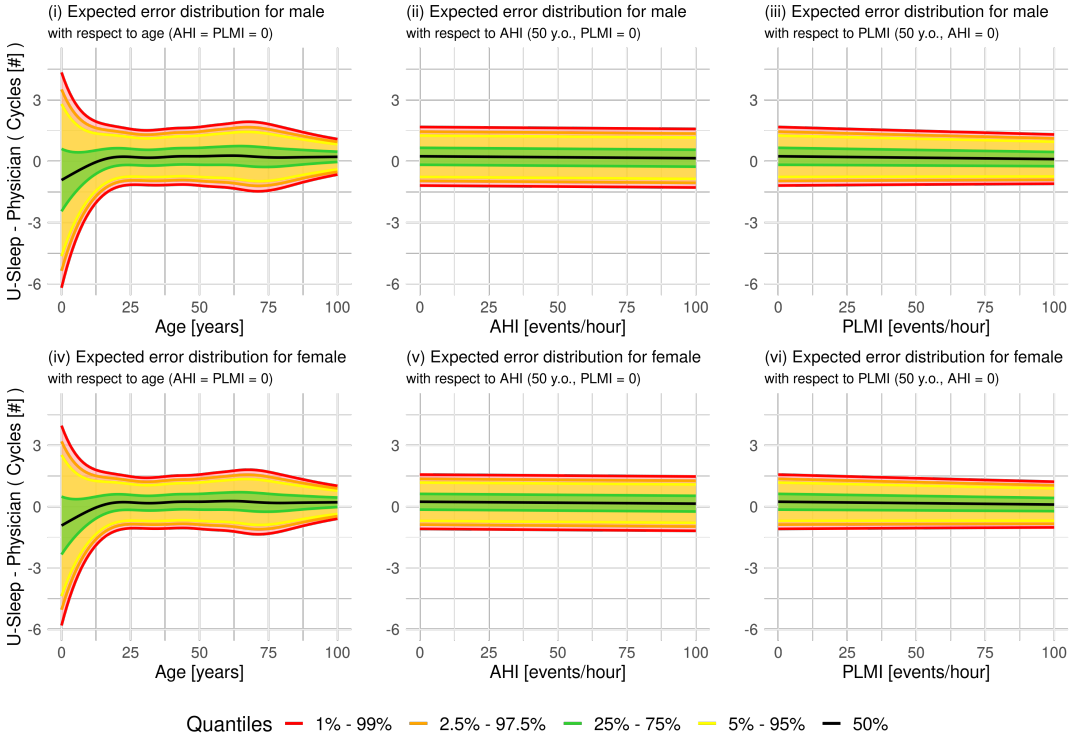
Notes: Expected distribution of the bias in the total sleep time (TST, minutes) based on the generalized normal distribution for U-Sleep predictions, stratified by gender (top row: males, bottom row: females). The graphs display the estimated distribution as a function of bias-inducing variables (age, AHI, and PLMI) on the horizontal axis. The solid black line represents the median, while the shaded areas correspond to different percentile ranges: 25-75% in green, 5-95% in yellow, 2.5-97.5% in orange, and 1-99% in red, illustrating the expected performance variability across subjects' characteristics.

Figure A.26: Expected distribution of the bias in the wake after sleep onset (WASO, minutes) for U-Sleep predictions.



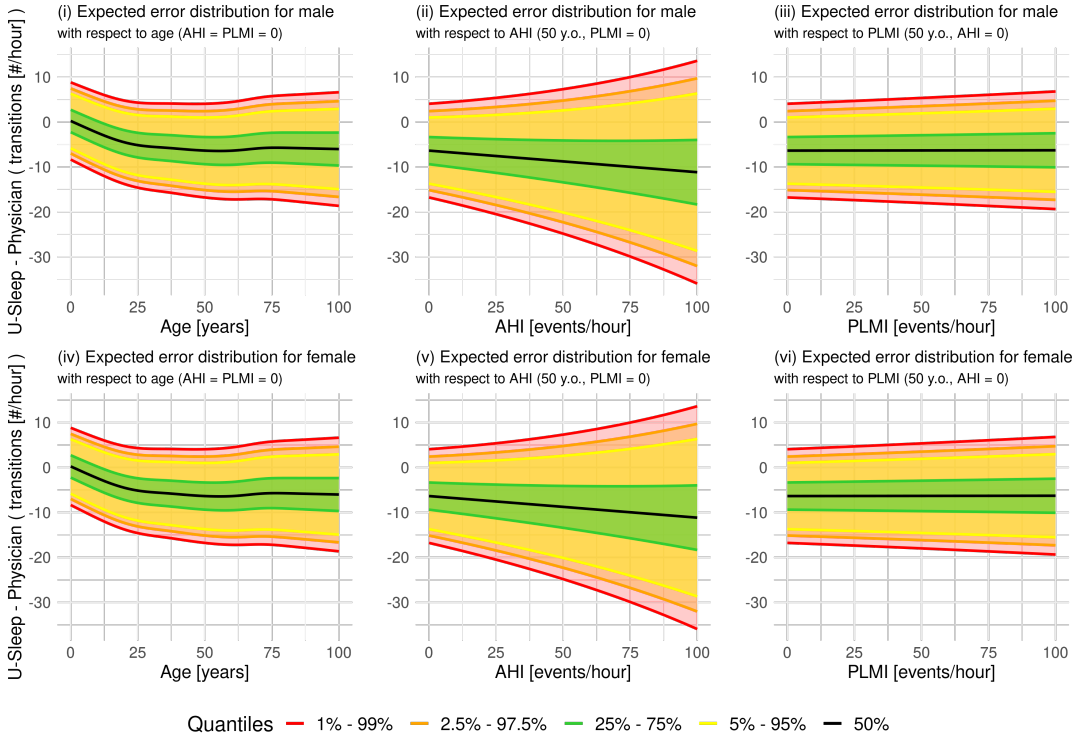
Notes: Expected distribution of the bias in the wake after sleep onset (WASO, minutes) based on the generalized normal distribution for U-Sleep predictions, stratified by gender (top row: males, bottom row: females). The graphs display the estimated distribution as a function of bias-inducing variables (age, AHI, and PLMI) on the horizontal axis. The solid black line represents the median, while the shaded areas correspond to different percentile ranges: 25-75% in green, 5-95% in yellow, 2.5-97.5% in orange, and 1-99% in red, illustrating the expected performance variability across subjects' characteristics.

Figure A.27: Expected distribution of the bias in the number (#) of sleep cycles for U-Sleep predictions.



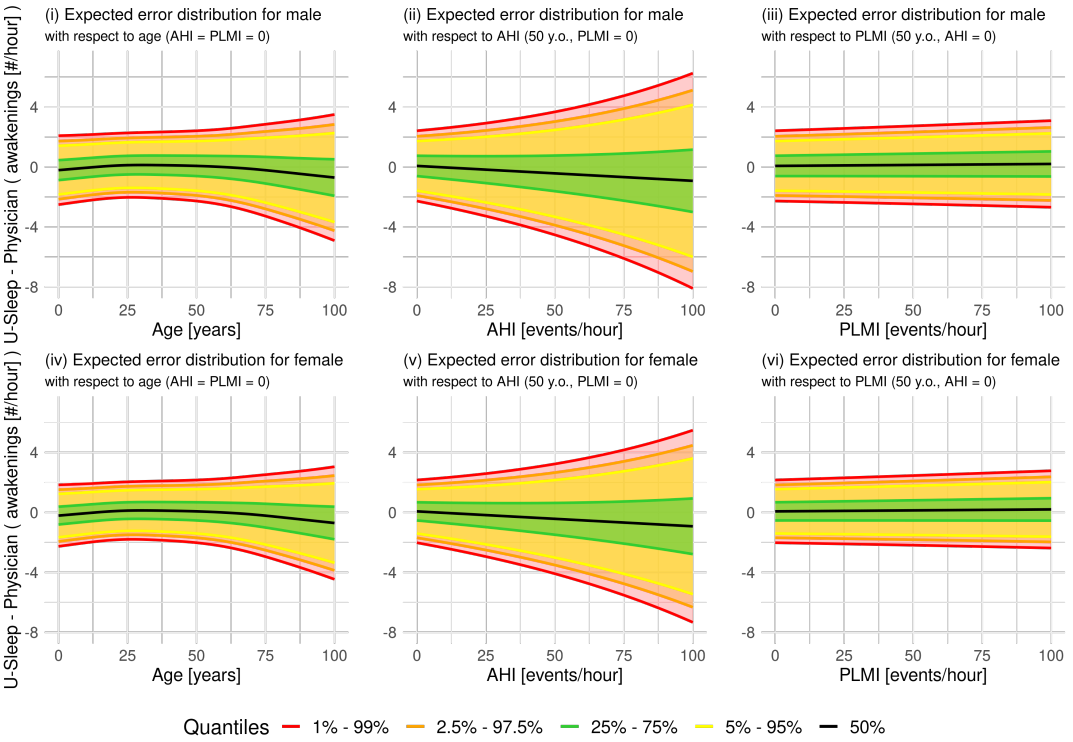
Notes: Expected distribution of the bias in the number (#) of sleep cycles based on the generalized normal distribution for U-Sleep predictions, stratified by gender (top row: males, bottom row: females). The graphs display the estimated distribution as a function of bias-inducing variables (age, AHI, and PLMI) on the horizontal axis. The solid black line represents the median, while the shaded areas correspond to different percentile ranges: 25-75% in green, 5-95% in yellow, 2.5-97.5% in orange, and 1-99% in red, illustrating the expected performance variability across subjects' characteristics.

Figure A.28: Expected distribution of the bias in the hourly rate (# / hour) of sleep stage transitions for U-Sleep predictions.



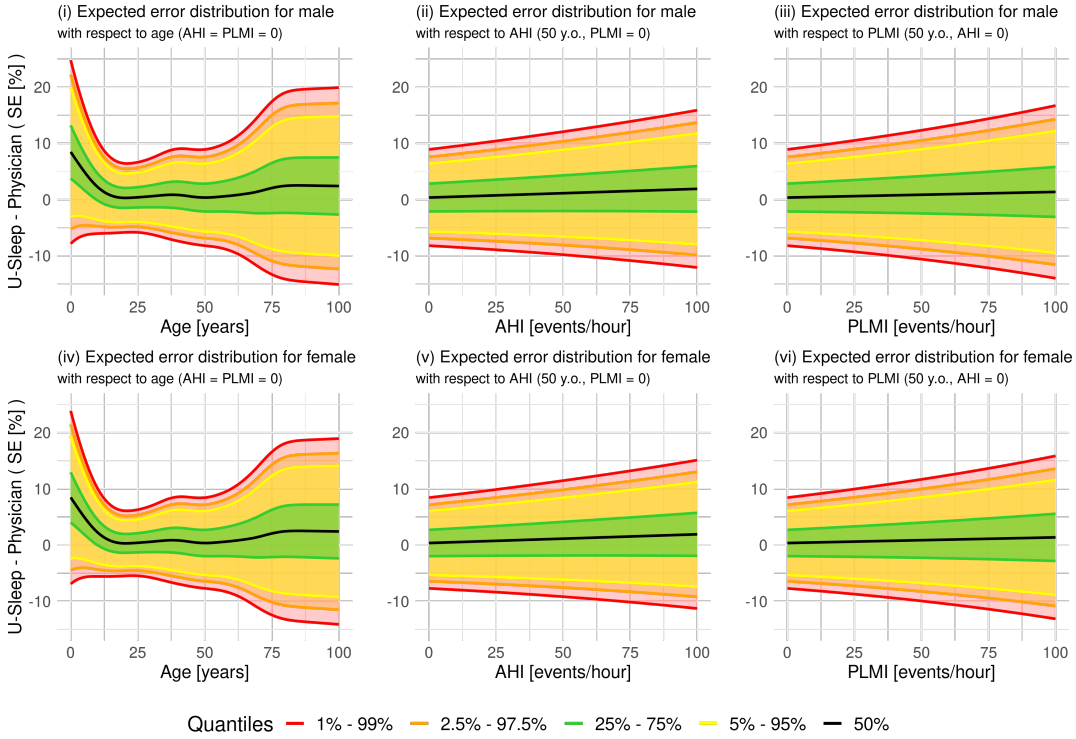
Notes: Expected distribution of the bias in the hourly rate (# / hour) of sleep stage transitions based on the generalized normal distribution for U-Sleep predictions, stratified by gender (top row: males, bottom row: females). The graphs display the estimated distribution as a function of bias-inducing variables (age, AHI, and PLMI) on the horizontal axis. The solid black line represents the median, while the shaded areas correspond to different percentile ranges: 25-75% in green, 5-95% in yellow, 2.5-97.5% in orange, and 1-99% in red, illustrating the expected performance variability across subjects' characteristics.

Figure A.29: Expected distribution of the bias in the hourly rate (# / hour) of awakenings for U-Sleep predictions.



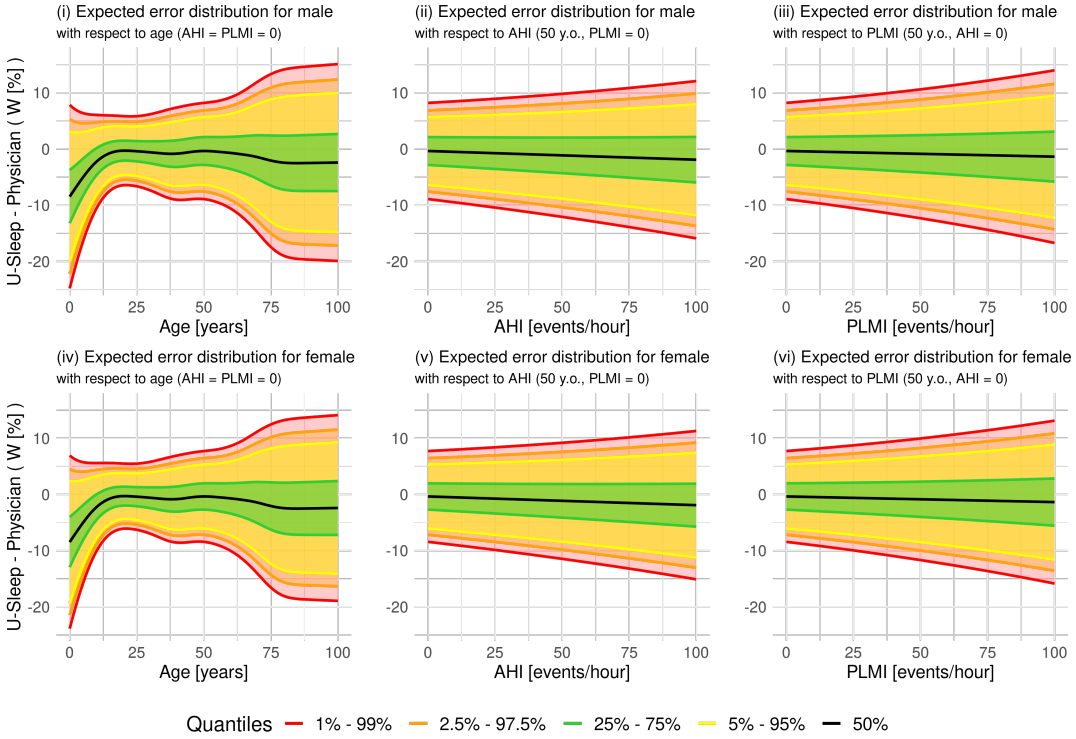
Notes: Expected distribution of the bias in the hourly rate (# / hour) of awakenings based on the generalized normal distribution for U-Sleep predictions, stratified by gender (top row: males, bottom row: females). The graphs display the estimated distribution as a function of bias-inducing variables (age, AHI, and PLMI) on the horizontal axis. The solid black line represents the median, while the shaded areas correspond to different percentile ranges: 25-75% in green, 5-95% in yellow, 2.5-97.5% in orange, and 1-99% in red, illustrating the expected performance variability across subjects' characteristics.

Figure A.30: Expected distribution of the bias in the sleep efficiency percentage (SE, %) for U-Sleep predictions.



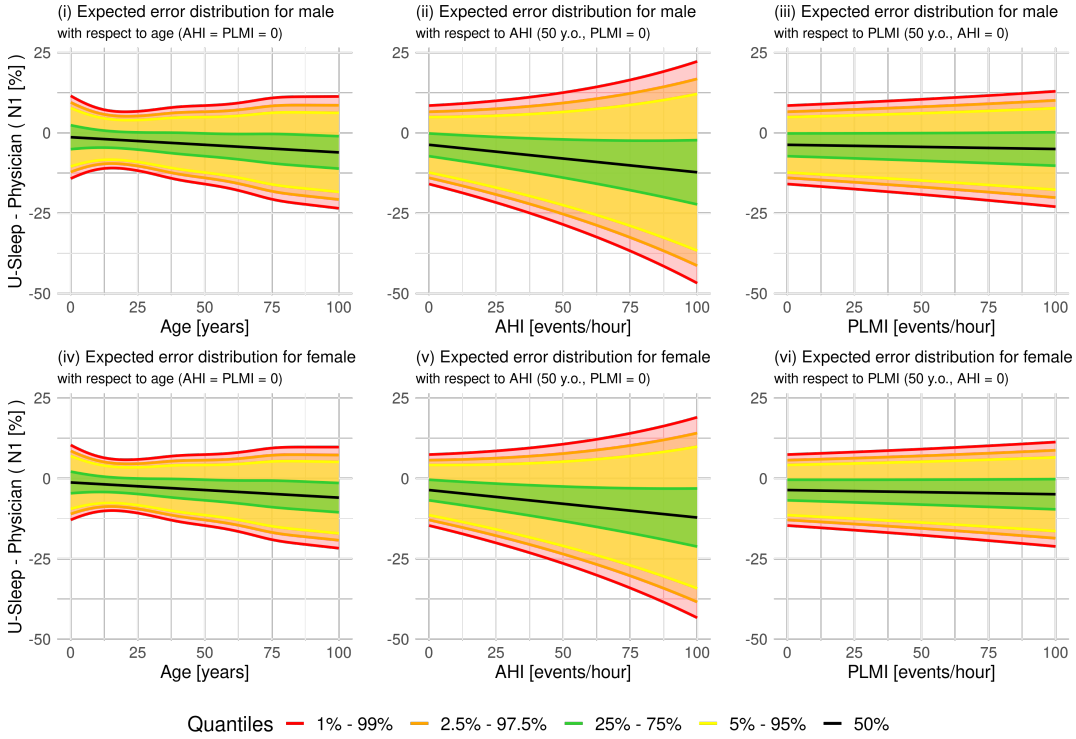
Notes: Expected distribution of the bias in the sleep efficiency percentage (SE, %) based on the generalized normal distribution for U-Sleep predictions, stratified by gender (top row: males, bottom row: females). The graphs display the estimated distribution as a function of bias-inducing variables (age, AHI, and PLMI) on the horizontal axis. The solid black line represents the median, while the shaded areas correspond to different percentile ranges: 25-75% in green, 5-95% in yellow, 2.5-97.5% in orange, and 1-99% in red, illustrating the expected performance variability across subjects' characteristics.

Figure A.31: Expected distribution of the bias in the wakefulness percentage after sleep onset (W, %) for U-Sleep predictions.



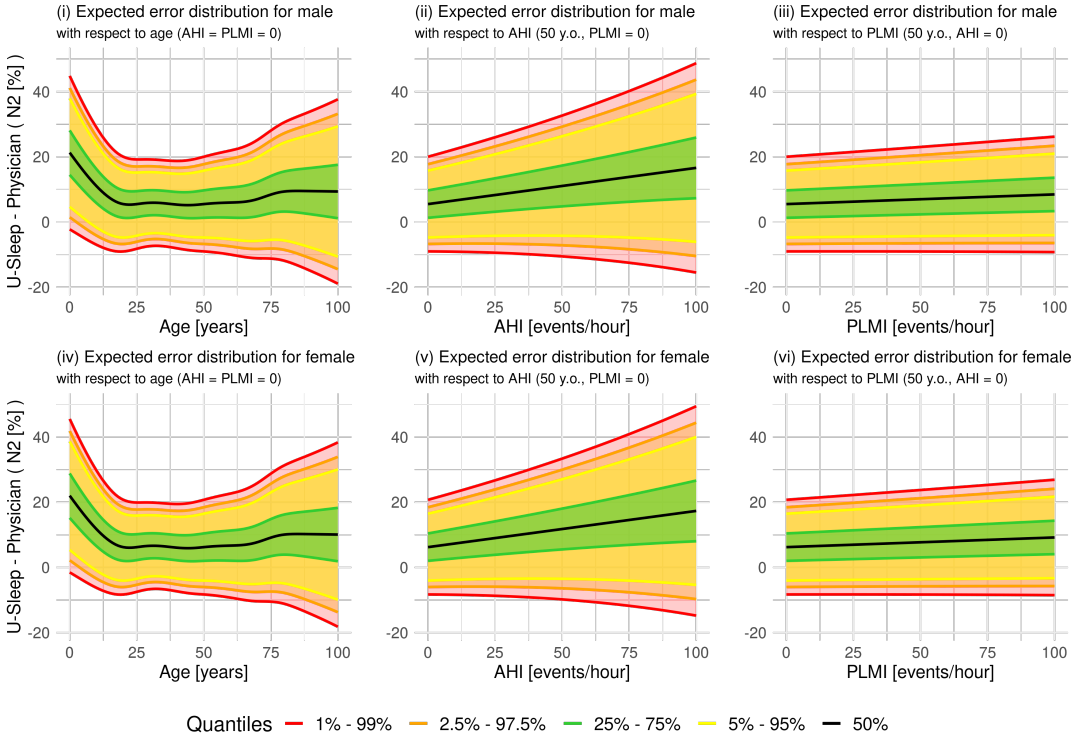
Notes: Expected distribution of the bias in the wakefulness percentage after sleep onset (W, %) based on the generalized normal distribution for U-Sleep predictions, stratified by gender (top row: males, bottom row: females). The graphs display the estimated distribution as a function of bias-inducing variables (age, AHI, and PLMI) on the horizontal axis. The solid black line represents the median, while the shaded areas correspond to different percentile ranges: 25-75% in green, 5-95% in yellow, 2.5-97.5% in orange, and 1-99% in red, illustrating the expected performance variability across subjects' characteristics.

Figure A.32: Expected distribution of the bias in the N1 sleep percentage after sleep onset (N1, %) for U-Sleep predictions.



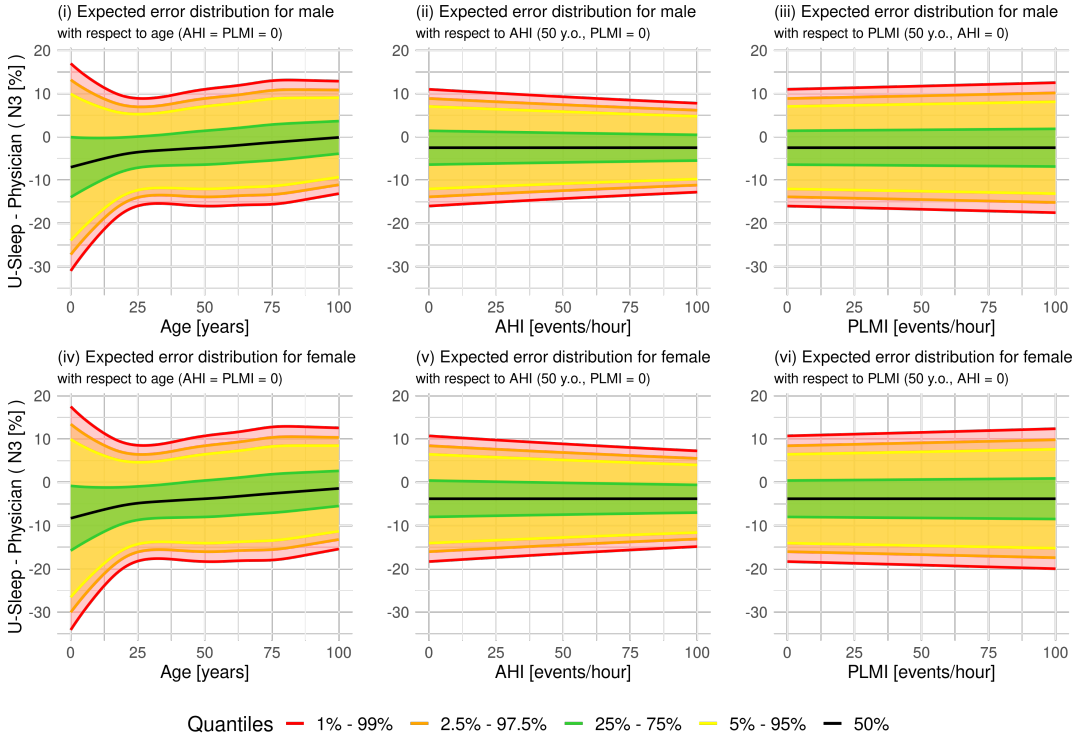
Notes: Expected distribution of the bias in the N1 sleep percentage after sleep onset (N1, %) based on the generalized normal distribution for U-Sleep predictions, stratified by gender (top row: males, bottom row: females). The graphs display the estimated distribution as a function of bias-inducing variables (age, AHI, and PLMI) on the horizontal axis. The solid black line represents the median, while the shaded areas correspond to different percentile ranges: 25-75% in green, 5-95% in yellow, 2.5-97.5% in orange, and 1-99% in red, illustrating the expected performance variability across subjects' characteristics.

Figure A.33: Expected distribution of the bias in the N2 sleep percentage after sleep onset (N2, %) for U-Sleep predictions.



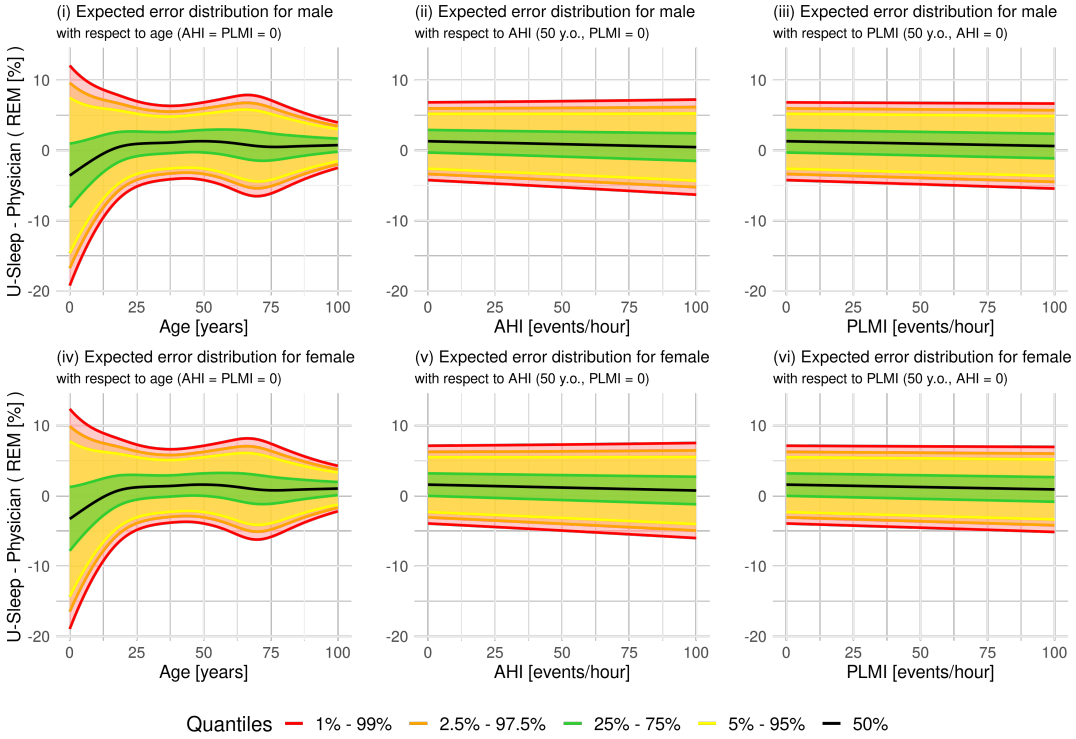
Notes: Expected distribution of the bias in the N2 sleep percentage after sleep onset (N2, %) based on the generalized normal distribution for U-Sleep predictions, stratified by gender (top row: males, bottom row: females). The graphs display the estimated distribution as a function of bias-inducing variables (age, AHI, and PLMI) on the horizontal axis. The solid black line represents the median, while the shaded areas correspond to different percentile ranges: 25-75% in green, 5-95% in yellow, 2.5-97.5% in orange, and 1-99% in red, illustrating the expected performance variability across subjects' characteristics.

Figure A.34: Expected distribution of the bias in the N3 sleep percentage after sleep onset (N3, %) for U-Sleep predictions.



Notes: Expected distribution of the bias in the N3 sleep percentage after sleep onset (N3, %) based on the generalized normal distribution for U-Sleep predictions, stratified by gender (top row: males, bottom row: females). The graphs display the estimated distribution as a function of bias-inducing variables (age, AHI, and PLMI) on the horizontal axis. The solid black line represents the median, while the shaded areas correspond to different percentile ranges: 25-75% in green, 5-95% in yellow, 2.5-97.5% in orange, and 1-99% in red, illustrating the expected performance variability across subjects' characteristics.

Figure A.35: Expected distribution of the bias in the REM sleep percentage after sleep onset (REM, %) for U-Sleep predictions.



Notes: Expected distribution of the bias in the REM sleep percentage after sleep onset (REM, %) based on the generalized normal distribution for U-Sleep predictions, stratified by gender (top row: males, bottom row: females). The graphs display the estimated distribution as a function of bias-inducing variables (age, AHI, and PLMI) on the horizontal axis. The solid black line represents the median, while the shaded areas correspond to different percentile ranges: 25-75% in green, 5-95% in yellow, 2.5-97.5% in orange, and 1-99% in red, illustrating the expected performance variability across subjects' characteristics.

Appendix B

Supplementary Materials for Chapter 5

B.1 Outcome model of Dirichlet regression

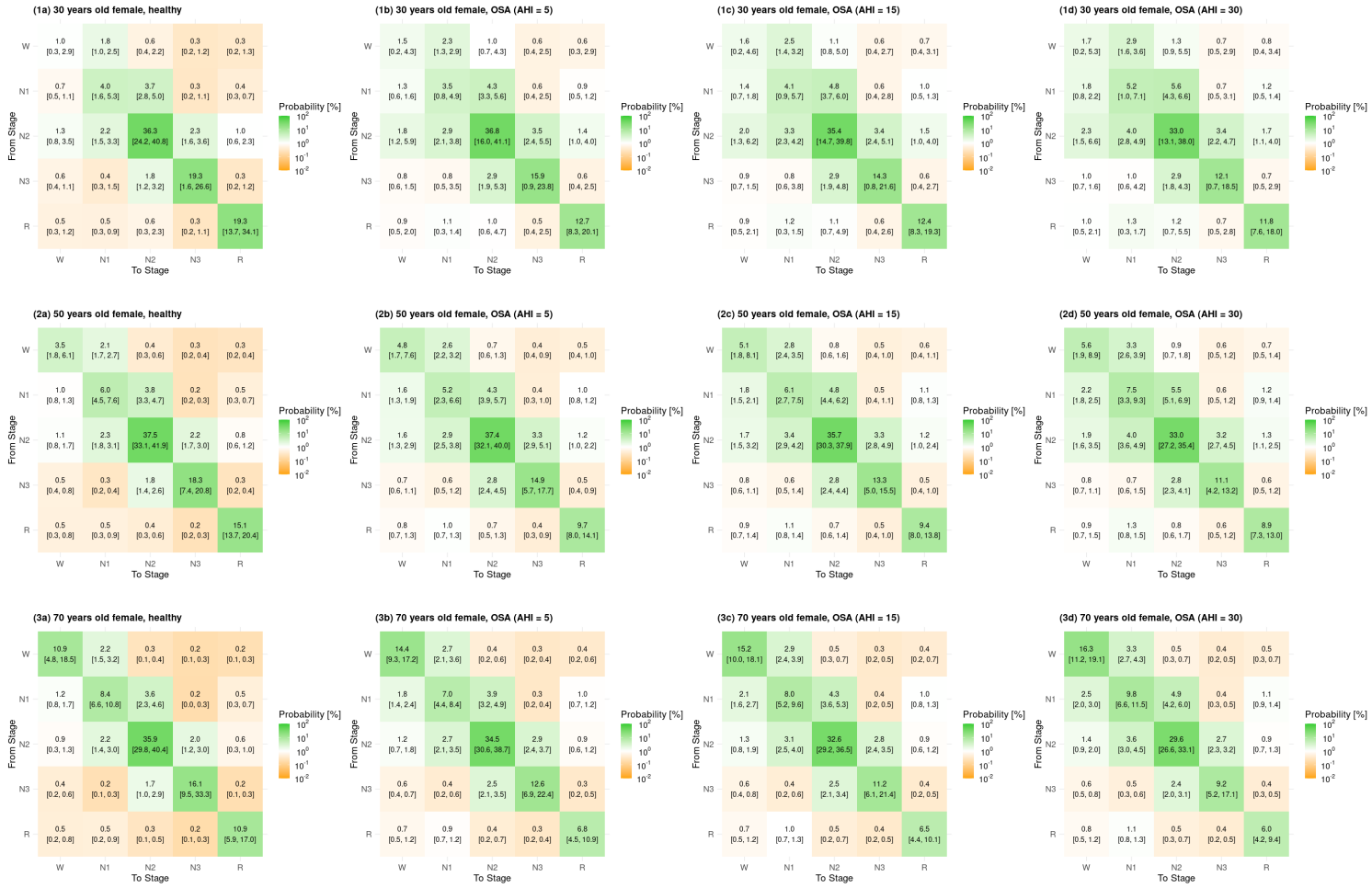
Table B.1: Estimated coefficients with bootstrapped 95% CI for the Dirichlet regression outcome model (Eq. 5.30).

α	Intercept	$\mathbb{I}_{\text{male}}$	$X_{(\text{Age}>50)/10}$	$\mathbb{I}_{\text{OSA}}$	$\mathbb{I}_{\text{OSA}} \times \mathbb{I}_{\text{male}}$	$\mathbb{I}_{\text{OSA}} \times X_{(\text{AHI}>5)/10}$	$\mathbb{I}_{\text{OSA}} \times \mathbb{I}_{\text{Insomnia\_Com}}$	$\mathbb{I}_{\text{OSA}} \times \mathbb{I}_{\text{NT1\_Com}}$	$\mathbb{I}_{\text{OSA}} \times \mathbb{I}_{\text{OtherHyp\_Com}}$	$\mathbb{I}_{\text{OSA}} \times \mathbb{I}_{\text{Parasomnia\_Com}}$	$\mathbb{I}_{\text{OSA}} \times \mathbb{I}_{\text{Movement\_Com}}$
W → W	**1.11*** (0.25, 1.94)	-0.18 (-0.64, 0.09)	0.00 (-0.10, 0.10)	**-0.59*** (-1.72, -0.19)	**0.50*** (0.12, 1.62)	0.12 (-1.42, 0.51)	-0.32 (-0.61, 0.36)	0.01 (-0.29, 0.22)	0.10 (-0.23, 1.68)	**-0.06*** (-0.11, -0.04)	**0.19*** (0.05, 0.44)
W → N1	0.65 (-0.33, 2.00)	0.17 (-0.59, 0.48)	**0.26*** (0.04, 0.65)	0.07 (-0.80, 0.48)	**0.17*** (0.05, 1.07)	0.01 (-0.04, 0.06)	**0.34*** (0.09, 0.65)	**-1.34*** (-1.47, -1.12)	**-0.96*** (-1.86, -0.57)	-0.13 (-0.31, 0.02)	**-0.14*** (-0.44, -0.01)
W → N2	**0.74*** (0.28, 2.03)	-0.02 (-0.04, 0.01)	**0.26*** (0.14, 0.48)	**1.18*** (0.80, 1.60)	0.01 (-0.73, 0.31)	-0.19 (-0.58, 0.06)	**-0.59*** (-1.23, -0.33)	-0.06 (-0.27, 0.59)	0.17 (-2.32, 0.92)	0.23 (-0.04, 0.81)	0.03 (-0.20, 0.23)
W → N3	-0.44 (-2.04, 0.32)	0.01 (-0.13, 0.18)	**-0.16*** (-0.61, -0.01)	-0.02 (-0.37, 1.51)	-0.40 (-1.64, 0.02)	0.04 (-0.28, 0.64)	-0.22 (-0.62, 0.09)	-0.00 (-0.05, 0.19)	**-0.23*** (-0.33, -0.12)	**0.33*** (0.17, 0.66)	**-1.59*** (-1.65, -1.50)
W → R	-0.49 (-2.21, 0.52)	0.02 (-0.18, 0.33)	-0.07 (-0.29, 0.08)	**0.14*** (0.03, 0.85)	**-0.04*** (-0.09, -0.00)	**0.48*** (0.22, 0.84)	-0.31 (-0.65, 0.05)	-0.12 (-0.37, 0.05)	**-0.67*** (-1.15, -0.07)	**-0.41*** (-1.18, -0.17)	-0.01 (-0.09, 0.40)
N1 → W	-0.06 (-0.11, 0.02)	**0.21*** (0.07, 0.34)	-0.18 (-0.57, 0.24)	**-0.62*** (-1.17, -0.32)	**-0.35*** (-0.59, -0.13)	**-0.71*** (-1.64, -0.34)	0.03 (-0.29, 1.21)	0.09 (-0.54, 0.30)	-0.27 (-0.89, 1.21)	-0.17 (-0.53, 0.06)	0.03 (-0.00, 0.23)
N1 → N1	0.31 (-0.28, 0.71)	**-0.16*** (-0.41, -0.01)	0.40 (-0.01, 1.66)	0.15 (-1.34, 0.54)	**0.48*** (0.18, 1.07)	0.05 (-0.55, 0.45)	0.03 (-0.03, 0.47)	**-0.03*** (-0.05, -0.01)	**1.03*** (0.49, 1.58)	**-0.82*** (-1.52, -0.30)	**-0.14*** (-0.26, -0.09)
N1 → N2	**0.82*** (0.39, 1.65)	-0.04 (-0.23, 0.15)	**0.26*** (0.14, 1.07)	-0.02 (-0.07, 0.03)	**0.28*** (0.06, 0.78)	**3.48*** (2.98, 3.98)	**-0.43*** (-0.86, -0.06)	-0.08 (-0.19, 0.04)	**-0.83*** (-2.01, -0.21)	0.36 (-0.19, 2.02)	0.01 (-0.46, 0.10)
N1 → N3	**0.51*** (0.13, 1.11)	**-1.52*** (-1.61, -1.39)	-0.25 (-1.03, 0.10)	-0.13 (-0.45, 0.12)	**-0.50*** (-1.44, -0.23)	-0.12 (-0.52, 1.39)	0.09 (-1.13, 0.44)	-0.09 (-0.24, 0.18)	-0.20 (-0.87, 0.37)	0.12 (-0.01, 1.19)	**-0.01*** (-0.03, -0.01)
N1 → R	**-0.87*** (-1.91, -0.26)	-0.06 (-0.18, 0.43)	-0.25 (-1.48, 0.18)	0.02 (-0.29, 0.68)	**-0.33*** (-0.82, -0.05)	0.15 (-0.04, 1.10)	**-0.07*** (-0.11, -0.04)	**0.13*** (0.03, 0.27)	**-1.47*** (-1.60, -1.34)	-0.02 (-0.87, 0.30)	-0.03 (-0.10, 0.02)
N2 → W	-0.07 (-0.72, 0.47)	0.02 (-0.01, 0.23)	0.01 (-0.03, 0.04)	**0.52*** (0.26, 0.83)	-0.06 (-0.49, 0.44)	**-0.76*** (-1.61, -0.47)	-0.11 (-0.33, 0.06)	-0.03 (-0.25, 0.12)	-0.12 (-0.23, 0.34)	-0.25 (-1.93, 0.25)	0.04 (-0.02, 0.22)
N2 → N1	**0.59*** (0.21, 1.00)	**-0.20*** (-0.38, -0.11)	0.09 (-0.12, 0.33)	**-0.76*** (-1.60, -0.36)	-0.03 (-0.45, 1.35)	0.19 (-1.56, 0.64)	-0.23 (-0.51, 0.28)	-0.02 (-0.19, 0.13)	**0.04*** (0.01, 0.24)	-0.04 (-0.10, 0.01)	**0.10*** (0.05, 0.22)
N2 → N2	0.18 (-0.18, 1.46)	0.07 (-0.46, 0.18)	**0.49*** (0.22, 1.03)	-0.04 (-0.65, 0.32)	0.06 (-0.03, 0.42)	**-0.17*** (-0.24, -0.09)	**0.34*** (0.18, 0.62)	**0.47*** (0.18, 0.84)	**-0.21*** (-0.36, -0.10)	**-0.28*** (-0.52, -0.02)	**-0.12*** (-0.33, -0.04)
N2 → N3	**0.20*** (0.10, 0.96)	**-0.01*** (-0.03, -0.01)	**0.31*** (0.10, 0.77)	**-1.54*** (-1.62, -1.42)	**-0.39*** (-0.91, -0.01)	-0.30 (-0.64, 0.02)	**-0.46*** (-0.93, -0.29)	-0.14 (-0.57, 1.46)	0.13 (-0.37, 0.25)	0.25 (-0.14, 0.90)	-0.03 (-0.16, 0.03)
N2 → R	**-0.54*** (-1.21, -0.18)	-0.03 (-0.11, 0.03)	**-0.51*** (-1.49, -0.25)	-0.05 (-0.13, 0.37)	0.08 (-1.43, 0.49)	-0.14 (-0.55, 0.63)	**-0.30*** (-0.61, -0.09)	**0.13*** (0.02, 0.77)	**-0.02*** (-0.04, -0.01)	**0.32*** (0.08, 0.95)	**2.57*** (2.29, 2.99)
N3 → W	-0.06 (-1.36, 0.32)	0.06 (-0.01, 0.25)	-0.04 (-0.87, 0.32)	0.03 (-0.01, 0.23)	**-0.04*** (-0.08, -0.00)	**0.52*** (0.22, 0.96)	**-0.80*** (-1.05, -0.44)	-0.32 (-0.79, 0.02)	-0.04 (-0.11, 0.02)	**-0.47*** (-1.47, -0.23)	-0.14 (-0.48, 1.26)
N3 → N1	-0.02 (-0.06, 0.02)	**0.12*** (0.07, 0.23)	**1.65*** (1.04, 2.06)	**-0.17*** (-0.32, -0.09)	0.11 (-0.08, 0.29)	**-0.93*** (-1.97, -0.39)	-0.15 (-0.55, 0.88)	0.05 (-1.78, 0.50)	0.03 (-0.04, 0.20)	-0.24 (-0.68, 0.10)	0.01 (-0.12, 0.57)
N3 → N2	0.08 (-0.16, 0.31)	**-0.13*** (-0.33, -0.07)	0.13 (-0.35, 1.50)	0.06 (-0.43, 0.15)	0.06 (-0.25, 0.60)	-0.30 (-1.07, 0.17)	0.07 (-0.05, 0.77)	**-0.13*** (-0.19, -0.08)	**0.11*** (0.04, 0.23)	**-0.98*** (-1.24, -0.68)	**-1.19*** (-1.97, -0.76)
N3 → N3	**0.36*** (0.10, 0.93)	-0.01 (-0.15, 0.09)	**0.33*** (0.13, 1.40)	**-0.01*** (-0.03, -0.00)	**0.36*** (0.14, 0.62)	**0.65*** (0.36, 1.00)	**-0.42*** (-0.89, -0.18)	-0.27 (-0.56, 0.01)	**-0.12*** (-0.34, -0.06)	0.02 (-0.31, 1.08)	0.28 (-1.23, 0.74)
N3 → R	**0.46*** (0.26, 0.85)	**-1.42*** (-1.58, -1.21)	**-0.90*** (-2.21, -0.52)	-0.04 (-0.10, 0.03)	**-0.39*** (-0.82, -0.17)	-0.16 (-0.55, 1.47)	0.03 (-1.16, 0.45)	-0.45 (-0.74, 0.28)	-0.02 (-0.16, 0.07)	-0.05 (-0.11, 0.18)	**-0.15*** (-0.22, -0.08)
R → W	**-0.55*** (-1.46, -0.30)	0.05 (-0.16, 0.91)	0.19 (-1.21, 0.69)	0.02 (-0.06, 0.18)	-0.12 (-0.47, 0.11)	**0.12*** (0.01, 0.78)	**-0.06*** (-0.10, -0.04)	**0.31*** (0.06, 0.62)	**-0.82*** (-1.10, -0.51)	**-0.31*** (-0.68, -0.04)	**-0.42*** (-0.92, -0.01)
R → N1	-0.05 (-0.67, 0.28)	0.03 (-0.04, 0.33)	0.03 (-0.02, 0.09)	**0.13*** (0.07, 0.24)	**0.68*** (0.29, 1.07)	**-0.35*** (-0.85, -0.03)	-0.09 (-0.28, 0.10)	**-0.60*** (-1.19, -0.32)	0.22 (-0.18, 1.54)	-0.01 (-1.06, 0.35)	0.08 (-0.62, 1.10)
R → N2	**-1.08*** (-1.32, -0.71)	**-0.19*** (-0.47, -0.05)	-0.12 (-0.46, 0.19)	**-0.10*** (-0.32, -0.02)	0.06 (-0.35, 1.66)	0.04 (-1.72, 0.47)	0.08 (-0.17, 0.49)	-0.24 (-0.67, 0.09)	**0.11*** (0.01, 0.73)	**-0.04*** (-0.07, -0.02)	**0.76*** (0.29, 1.16)
R → N3	-0.21 (-0.52, 0.51)	-0.04 (-0.89, 0.19)	**0.74*** (0.41, 1.41)	0.01 (-0.12, 0.12)	**0.14*** (0.01, 0.85)	**-0.13*** (-0.19, -0.09)	**0.31*** (0.14, 0.68)	**2.76*** (1.65, 3.17)	-0.26 (-0.72, 0.03)	-0.08 (-0.24, 0.04)	**-0.68*** (-1.48, -0.13)
R → R	-0.04 (-0.11, 0.16)	0.00 (-0.02, 0.02)	**0.50*** (0.13, 0.99)	**-0.91*** (-1.29, -0.58)	**-0.50*** (-1.05, -0.16)	-0.26 (-0.55, 0.01)	**-0.25*** (-1.11, -0.05)	-0.46 (-1.10, 1.20)	-0.12 (-1.42, 0.25)	-0.08 (-0.25, 0.26)	-0.19 (-1.05, 0.31)

Notes: Significant estimates are highlighted in bold (*). Rows correspond to individual dimensions representing each of the 25 possible sleep-stage transition proportions.

B.2 Comparison based on matrices of transition proportions P

Figure B.1: Expected matrices of transition proportions **P** for healthy females and females with different OSA severities, each stratified by age.



Notes: OSA severities are represented by AHI levels of 5, 15, and 30, and ages by 30, 50, and 70 years. Estimates are shown with 95% bootstrapped confidence intervals (CIs).

Figure B.2: Differences (CATE) in matrices of transition proportions **P** between healthy females and females with different OSA severities, each stratified by age.

Notes: OSA severities are represented by AHI levels of 5, 15, and 30, and ages by 30, 50, and 70 years. Estimates are shown with 95% bootstrapped confidence intervals (CIs).

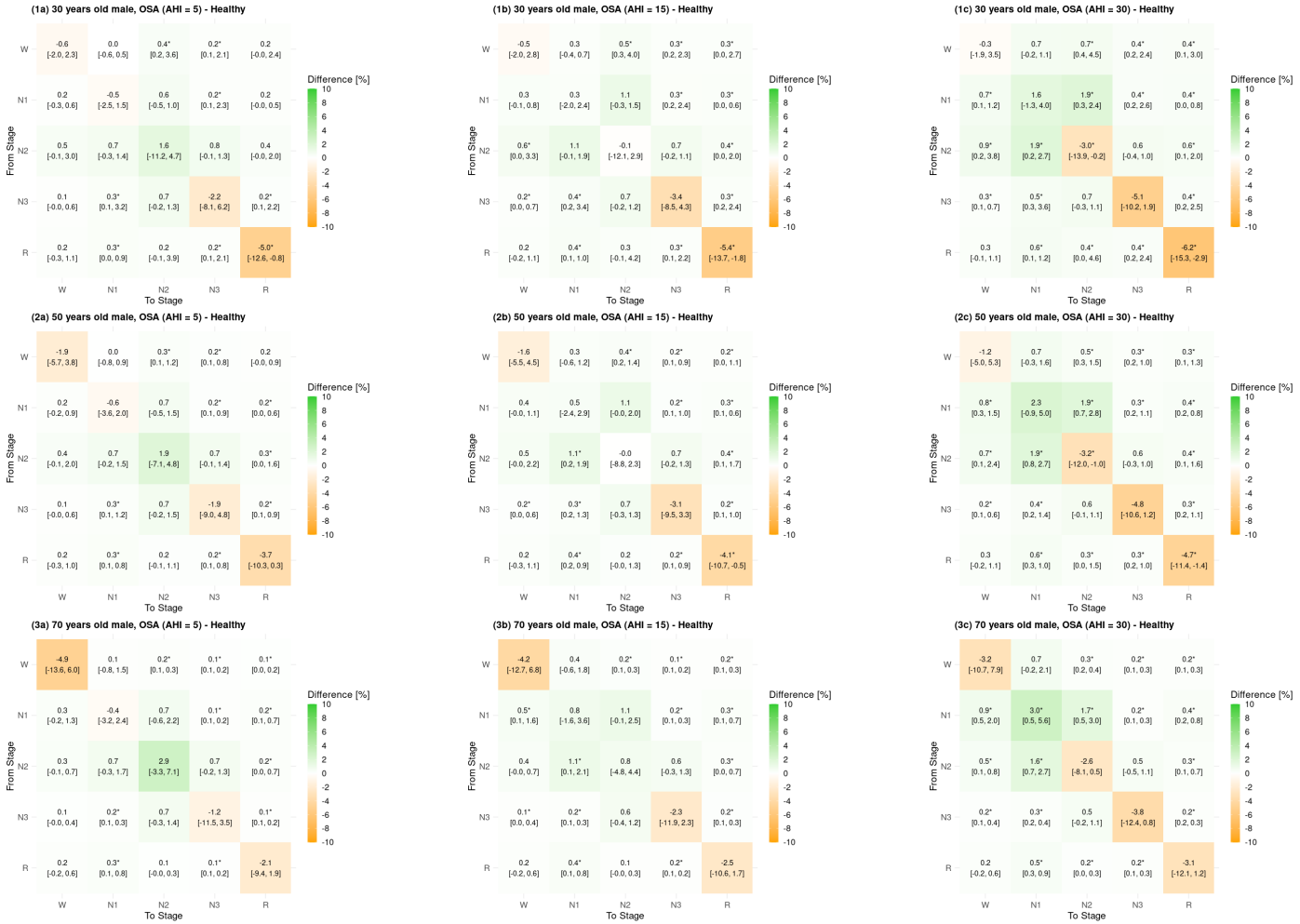
Figure B.3: Risk ratio (RR-CATE) of matrices of transition proportions **P** between healthy females and females with different OSA severities, each stratified by age.

Notes: OSA severities are represented by AHI levels of 5, 15, and 30, and ages by 30, 50, and 70 years. Estimates are shown with 95% bootstrapped confidence intervals (CIs).

Figure B.4: Expected matrices of transition proportions **P** for healthy males and males with different OSA severities, each stratified by age.

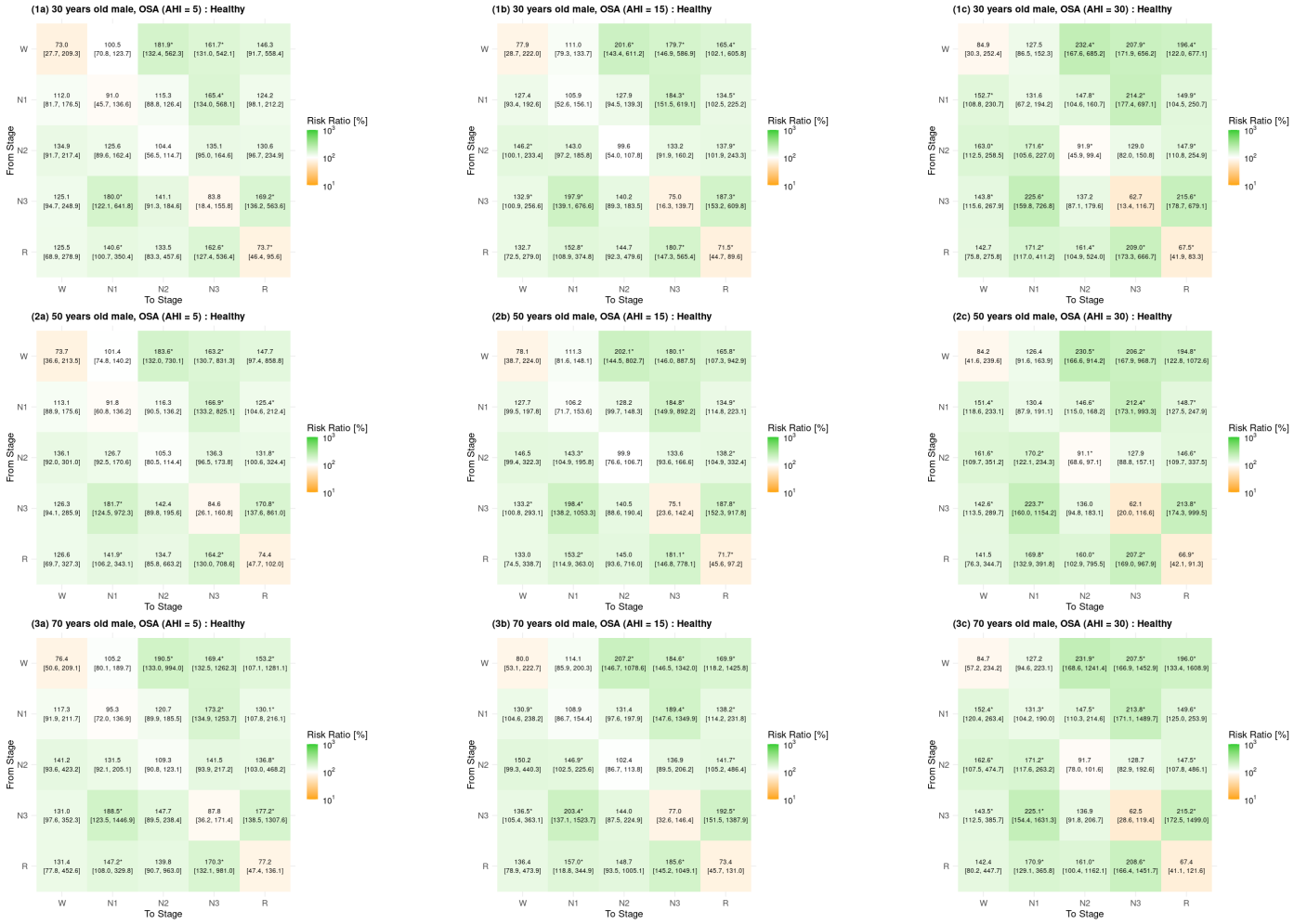
Notes: OSA severities are represented by AHI levels of 5, 15, and 30, and ages by 30, 50, and 70 years. Estimates are shown with 95% bootstrapped confidence intervals (CIs).

Figure B.5: Differences (CATE) in matrices of transition proportions **P** between healthy males and males with different OSA severities, each stratified by age.



Notes: OSA severities are represented by AHI levels of 5, 15, and 30, and ages by 30, 50, and 70 years. Estimates are shown with 95% bootstrapped confidence intervals (CIs).

Figure B.6: Risk ratio (RR-CATE) of matrices of transition proportions **P** between healthy males and males with different OSA severities, each stratified by age.



Notes: OSA severities are represented by AHI levels of 5, 15, and 30, and ages by 30, 50, and 70 years. Estimates are shown with 95% bootstrapped confidence intervals (CIs).

B.3 Effect tables for markers of sleep macro-structure and dynamics

Table B.2: Expected probabilities and estimated OSA effects (CATE, RR-CATE) for 30-year-old females.

Quantity	Estimate	Healthy	O1: OSA (AHI = 5)	O2: OSA (AHI = 15)	O3: OSA (AHI = 30)
P(W)	%	4.17 (2.85, 7.15)	6.28 (4.45, 11.24)	6.85 (4.75, 11.67)	7.74 (5.34, 12.55)
	CATE		**2.1*** (0.87, 4.29)	**2.68*** (1.42, 4.94)	**3.57*** (2.11, 5.91)
	RR-CATE		**150.47*** (118.71, 199.74)	**164.35*** (129.93, 212.12)	**185.63*** (143.44, 236.11)
P(N1)	%	8.93 (7.14, 11.04)	10.52 (8.87, 12.22)	12 (10.17, 13.57)	14.45 (11.34, 16.46)
	CATE		1.59 (-0.52, 3.74)	**3.07*** (0.78, 5.23)	**5.53*** (2.8, 7.61)
	RR-CATE		117.87 (94.82, 146.85)	**134.41*** (107.51, 164.96)	**161.94*** (125.03, 197.44)
P(N2)	%	43.03 (35.97, 46.7)	45.99 (35.57, 49.56)	45.29 (34.95, 48.5)	43.98 (34.44, 47.83)
	CATE		2.95 (-2.47, 7.24)	2.26 (-3.24, 6.9)	0.94 (-4.71, 5.28)
	RR-CATE		106.86 (94.35, 118.57)	105.25 (92.97, 117.03)	102.2 (89.66, 113.11)
P(N3)	%	22.45 (7.76, 29.61)	21.02 (11.32, 28.36)	19.62 (11.16, 26.19)	17.65 (11.24, 23.86)
	CATE		-1.43 (-6.72, 5.62)	-2.83 (-7.76, 5.96)	-4.8 (-9.49, 6.39)
	RR-CATE		93.63 (75.77, 164.23)	87.4 (71.79, 165.62)	78.61 (63.55, 171.61)
P(REM)	%	21.42 (15.78, 39.37)	16.2 (11.82, 28.76)	16.24 (11.7, 28.3)	16.18 (11.55, 27.31)
	CATE		**-5.22*** (-12.13, -1.35)	**-5.18*** (-12.89, -1.7)	**-5.24*** (-13.99, -1.88)
	RR-CATE		**75.62*** (63.08, 93.18)	**75.8*** (63.54, 91.42)	**75.53*** (62.73, 90.04)
P((N1,N2,N3,REM) $\rightarrow$ W)	%	3.12 (2.2, 6.87)	4.82 (3.24, 10.16)	5.29 (3.6, 10.48)	6.02 (4.04, 11.3)
	CATE		**1.69*** (0.77, 3.64)	**2.16*** (1.12, 3.83)	**2.89*** (1.6, 4.53)
	RR-CATE		**154.24*** (126.08, 195.75)	**169.25*** (135.59, 209.54)	**192.55*** (151.32, 230.47)
P((N1,N2) $\rightarrow$ W)	%	2.01 (1.35, 4.25)	3.11 (2.1, 7.13)	3.46 (2.36, 7.37)	4.02 (2.72, 7.83)
	CATE		**1.1*** (0.46, 2.38)	**1.45*** (0.74, 2.88)	**2.01*** (1.18, 3.49)
	RR-CATE		**154.75*** (121.62, 200.22)	**172.33*** (137.19, 215.64)	**200.35*** (154.77, 253.65)
P(N3 $\rightarrow$ W)	%	0.59 (0.45, 1.09)	0.84 (0.61, 1.46)	0.89 (0.66, 1.49)	0.98 (0.72, 1.56)
	CATE		**0.25*** (0.03, 0.53)	**0.3*** (0.1, 0.58)	**0.39*** (0.17, 0.69)
	RR-CATE		**141.77*** (104.75, 190.72)	**151.7*** (113.57, 199.84)	**165.96*** (124.65, 219.4)
P(REM $\rightarrow$ W)	%	0.53 (0.25, 1.21)	0.88 (0.46, 2.05)	0.93 (0.5, 2.09)	1.02 (0.51, 2.12)
	CATE		**0.35*** (0.08, 1.1)	**0.41*** (0.12, 0.93)	**0.49*** (0.16, 0.96)
	RR-CATE		**166.23*** (115.25, 259.61)	**177.15*** (122.92, 270.8)	**192.6*** (130.02, 293.61)
P(NREM $\rightleftarrows$ REM)	%	3.19 (2.19, 7.53)	5.53 (4.07, 15.36)	6.02 (4.39, 16.38)	6.78 (4.98, 17.83)
	CATE		**2.33*** (1.19, 7.93)	**2.83*** (1.56, 8.93)	**3.59*** (2.06, 10.38)
	RR-CATE		**173.18*** (135.99, 244.33)	**188.79*** (145.45, 260.22)	**212.48*** (159.99, 292.82)
P(N1 $\rightleftarrows$ N2)	%	5.89 (4.25, 8.34)	7.17 (5.41, 9.46)	8.1 (6.07, 9.92)	9.61 (7.14, 11.37)
	CATE		1.28 (-0.2, 2.42)	**2.21*** (0.59, 3.37)	**3.72*** (1.57, 5.04)
	RR-CATE		121.73 (97.19, 150.33)	**137.52*** (108.78, 167.48)	**163.24*** (125.05, 193.52)
P(Sleep compactness)	%	92.82 (85.85, 95.05)	89.31 (76.67, 92.32)	88.21 (75.49, 91.52)	86.49 (73.31, 90.45)
	CATE		**-3.52*** (-8.83, -1.43)	**-4.62*** (-10.46, -2.37)	**-6.34*** (-12.18, -4.02)
	RR-CATE		**96.21*** (89.76, 98.45)	**95.02*** (87.88, 97.4)	**93.17*** (85.11, 95.75)
P(Sleep fragmentation)	%	6.13 (4.26, 13.87)	9.24 (6.28, 23.18)	10.23 (7.14, 24.35)	11.79 (8.03, 26.05)
	CATE		**3.11*** (1.43, 9.04)	**4.1*** (2.22, 10.58)	**5.66*** (3.38, 11.96)
	RR-CATE		**150.7*** (124.39, 190.2)	**166.87*** (139.17, 204.44)	**192.33*** (160.07, 236.24)
P(Sleep-stage compactness)	%	78.91 (61.89, 82.74)	68.93 (36.64, 74.62)	66.28 (34.47, 72.21)	62.15 (30.04, 68.79)
	CATE		**-9.98*** (-25.15, -6.58)	**-12.64*** (-28.1, -9.02)	**-16.76*** (-32.37, -12.66)
	RR-CATE		**87.35*** (58.8, 91.84)	**83.98*** (55.19, 88.7)	**78.76*** (49.17, 84.41)
P(Sleep-stage fragmentation)	%	13.91 (10.74, 23.49)	20.37 (16, 40.05)	21.93 (17.32, 40.28)	24.33 (18.83, 42.09)
	CATE		**6.46*** (3.09, 16.83)	**8.02*** (4.56, 17.45)	**10.42*** (6.5, 18.44)
	RR-CATE		**146.46*** (124.33, 187.22)	**157.66*** (134, 198.38)	**174.94*** (148.28, 216.31)
P(W $\rightarrow$ W)	%	1.05 (0.33, 2.92)	1.46 (0.18, 4.27)	1.57 (0.19, 4.64)	1.73 (0.17, 5.32)
	CATE		0.41 (-0.35, 2.23)	0.52 (-0.28, 2.52)	0.68 (-0.21, 2.97)
	RR-CATE		139.22 (42.54, 274.01)	149.7 (44.75, 299.99)	164.95 (45.63, 333.06)
P(N1 $\rightarrow$ N1)	%	4.01 (1.59, 5.26)	3.53 (0.77, 4.91)	4.14 (0.86, 5.73)	5.2 (1.03, 7.1)
	CATE		-0.48 (-1.85, 0.73)	0.13 (-1.44, 1.44)	1.19 (-1.11, 2.86)
	RR-CATE		88.06 (37.68, 118.66)	103.3 (43.8, 136.66)	129.7 (54.59, 171.73)
P(N2 $\rightarrow$ N2)	%	36.34 (24.17, 40.75)	36.85 (16.05, 41.1)	35.44 (14.74, 39.78)	33.03 (13.14, 38)
	CATE		0.5 (-6.46, 5.55)	-0.9 (-8.42, 4.06)	-3.31 (-10.99, 1.39)
	RR-CATE		101.39 (68.82, 115.48)	97.51 (62.45, 110.85)	90.89 (54.45, 104.41)
P(N3 $\rightarrow$ N3)	%	19.26 (1.59, 26.61)	15.9 (0.87, 23.85)	14.33 (0.78, 21.61)	12.11 (0.68, 18.55)
	CATE		-3.36 (-8.53, 0.35)	**-4.93*** (-9.66, -0.44)	**-7.15*** (-11.67, -0.76)
	RR-CATE		82.55 (43.2, 104.4)	**74.38*** (40.56, 91.4)	**62.87*** (29.6, 77.92)
P(REM $\rightarrow$ REM)	%	19.3 (13.72, 34.14)	12.66 (8.28, 20.06)	12.37 (8.32, 19.29)	11.81 (7.62, 18.04)
	CATE		**-6.65*** (-16.35, -2.47)	**-6.93*** (-17.11, -2.97)	**-7.49*** (-18.28, -3.86)
	RR-CATE		**65.57*** (49.07, 83.92)	**64.08*** (46.32, 81.8)	**61.18*** (42.31, 77.66)
P(W-fragmentation)	%	3.01 (2.09, 7.01)	4.42 (3.06, 12.3)	4.94 (3.51, 13.03)	5.77 (4, 14.79)
	CATE		**1.41*** (0.65, 5.38)	**1.94*** (1.05, 6.47)	**2.77*** (1.74, 7.97)
	RR-CATE		**147.02*** (125.01, 186.42)	**164.39*** (139.85, 201.78)	**192.1*** (163.44, 238.83)
P(N1-fragmentation)	%	5.16 (3.94, 7.31)	7.02 (5.7, 9.56)	7.86 (6.47, 10.26)	9.21 (7.52, 11.44)
	CATE		**1.86*** (0.96, 2.8)	**2.71*** (1.65, 3.78)	**4.05*** (2.86, 5.17)
	RR-CATE		**136.16*** (115.95, 162.61)	**152.47*** (130.35, 178.28)	**178.61*** (143.53, 209.53)
P(N2-fragmentation)	%	6.78 (5.13, 12.06)	9.65 (7.38, 18.19)	10.31 (7.8, 18.48)	11.33 (8.47, 18.81)
	CATE		**2.86*** (1.26, 5.9)	**3.53*** (1.81, 5.97)	**4.55*** (2.69, 6.7)
	RR-CATE		**142.19*** (118.74, 170.61)	**151.99*** (127.54, 182.91)	**167.08*** (139.25, 199.61)
P(N3-fragmentation)	%	3.15 (2.3, 6.58)	5.05 (3.71, 12.17)	5.26 (3.92, 12.5)	5.54 (4.11, 12.97)
	CATE		**1.9*** (0.77, 6.35)	**2.11*** (1.02, 6.56)	**2.39*** (1.18, 6.62)
	RR-CATE		**160.31*** (123.48, 224.18)	**166.87*** (132.04, 230.58)	**175.86*** (138.11, 238.99)
P(REM-fragmentation)	%	1.94 (1.21, 4.89)	3.47 (2.45, 10.09)	3.79 (2.7, 10.78)	4.26 (3.06, 11.58)
	CATE		**1.53*** (0.75, 5.22)	**1.84*** (0.97, 5.95)	**2.32*** (1.29, 7)
	RR-CATE		**178.76*** (134.92, 258.28)	**194.95*** (147.82, 280.14)	**219.51*** (165.81, 312.96)

Notes: Probabilities are expressed as percentages. Estimates include conditional average treatment effects (CATE) and risk-ratio CATE (RR-CATE).

Table B.3: Expected probabilities and estimated OSA effects (CATE, RR-CATE) for 50-year-old females.

Quantity	Estimate	Healthy	O1: OSA (AHI = 5)	O2: OSA (AHI = 15)	O3: OSA (AHI = 30)
P(W)	%	6.59 (4.83, 9.19)	9.51 (7.29, 12.83)	10.3 (8.11, 13.63)	11.46 (9.19, 14.72)
	CATE		**2.92*** (0.29, 6.05)	**3.7*** (1.11, 6.77)	**4.87*** (2.14, 7.86)
	RR-CATE		**144.3*** (104.48, 208.14)	**156.17*** (114.78, 221.8)	**173.85*** (126.1, 244.19)
P(N1)	%	11.2 (9.66, 13.59)	12.33 (10.63, 13.97)	14.04 (12.07, 15.91)	16.85 (14.28, 18.41)
	CATE		1.13 (-1.63, 3.67)	2.83 (-0.11, 5.36)	**5.65*** (2.13, 7.98)
	RR-CATE		110.1 (86.24, 135.03)	125.29 (99.12, 151.07)	**150.4*** (117.31, 180.77)
P(N2)	%	43.99 (40.42, 48.3)	45.87 (43.03, 49)	44.83 (42.38, 47.65)	43 (40.43, 45.75)
	CATE		1.88 (-3.03, 6.14)	0.85 (-4.33, 5.13)	-0.99 (-6.38, 3.08)
	RR-CATE		104.28 (93.63, 115.16)	101.93 (91, 112.67)	97.75 (86.82, 107.5)
P(N3)	%	21.29 (11.26, 24.02)	19.49 (12.84, 22.54)	18.05 (12.38, 20.5)	16.02 (11.57, 18.25)
	CATE		-1.81 (-5.97, 3.33)	-3.25 (-7.51, 2.15)	-5.28 (-9.29, 1.1)
	RR-CATE		91.5 (72.53, 128.84)	84.75 (68.01, 123.55)	75.22 (60.11, 112.02)
P(REM)	%	16.92 (15.51, 22.66)	12.8 (11.05, 18.45)	12.79 (11.08, 18.39)	12.67 (11.04, 18.38)
	CATE		**-4.13*** (-7.41, -1.08)	**-4.14*** (-7.11, -1.2)	**-4.25*** (-7.32, -1.23)
	RR-CATE		**75.61*** (63.48, 93.29)	**75.56*** (64.48, 92.93)	**74.9*** (63.2, 92.36)
P((N1,N2,N3,REM) $\to$ W)	%	3.08 (2.66, 4.38)	4.71 (4.06, 7.35)	5.16 (4.48, 7.77)	5.86 (5.02, 8.6)
	CATE		**1.63*** (0.91, 3.19)	**2.08*** (1.34, 3.81)	**2.78*** (1.94, 4.76)
	RR-CATE		**153.03*** (129.36, 197.2)	**167.68*** (141.27, 208.08)	**190.25*** (158.06, 239.26)
P((N1,N2) $\to$ W)	%	2.05 (1.65, 2.95)	3.16 (2.72, 4.69)	3.52 (3.05, 5.1)	4.08 (3.47, 5.9)
	CATE		**1.11*** (0.59, 1.99)	**1.46*** (0.95, 2.46)	**2.03*** (1.44, 3.24)
	RR-CATE		**154.05*** (125.49, 200.21)	**171.38*** (141.51, 224.97)	**198.82*** (162.16, 251.85)
P(N3 $\to$ W)	%	0.52 (0.42, 0.77)	0.72 (0.57, 1.09)	0.77 (0.61, 1.12)	0.83 (0.67, 1.15)
	CATE		**0.2*** (0.03, 0.46)	**0.25*** (0.07, 0.49)	**0.31*** (0.14, 0.54)
	RR-CATE		**139.15*** (105.33, 191.21)	**148.04*** (113.09, 200.2)	**160.43*** (122.7, 213.34)
P(REM $\to$ W)	%	0.51 (0.33, 0.79)	0.83 (0.66, 1.34)	0.88 (0.7, 1.4)	0.94 (0.74, 1.52)
	CATE		**0.32*** (0.09, 0.78)	**0.37*** (0.13, 0.77)	**0.44*** (0.17, 0.84)
	RR-CATE		**163.17*** (114.04, 284.16)	**172.87*** (121.72, 302.79)	**186.19*** (128.28, 307.05)
P(NREM $\rightleftarrows$ REM)	%	2.75 (2.07, 3.73)	4.77 (3.96, 7.34)	5.17 (4.34, 8)	5.76 (4.85, 8.8)
	CATE		**2.01*** (1.08, 4.42)	**2.41*** (1.34, 5.09)	**3.01*** (1.78, 6.07)
	RR-CATE		**173.2*** (133.27, 253.24)	**187.76*** (140.87, 274.42)	**209.38*** (153.92, 308.82)
P(N1 $\rightleftarrows$ N2)	%	6.06 (5.12, 7.63)	7.24 (6.43, 9.46)	8.13 (7.32, 10.42)	9.56 (8.73, 11.75)
	CATE		1.18 (-0.08, 2.65)	**2.07*** (0.77, 3.61)	**3.5*** (2.11, 5.21)
	RR-CATE		119.5 (98.73, 150.25)	**134.22*** (111.38, 169.78)	**157.84*** (130.84, 192.88)
P(Sleep compactness)	%	90.39 (87.53, 92.32)	86.29 (81.86, 88.51)	85.04 (80.68, 87.24)	83.15 (78.95, 85.5)
	CATE		**-4.1*** (-8.04, -1.22)	**-5.35*** (-8.92, -2.57)	**-7.24*** (-11.29, -4.49)
	RR-CATE		**95.46*** (91.07, 98.63)	**94.08*** (89.48, 97.11)	**91.99*** (87.49, 94.94)
P(Sleep fragmentation)	%	6.09 (5.31, 8.39)	8.91 (7.69, 14.29)	9.82 (8.49, 15.36)	11.25 (9.61, 16.91)
	CATE		**2.82*** (1.55, 6.03)	**3.73*** (2.34, 7.29)	**5.16*** (3.61, 9.23)
	RR-CATE		**146.23*** (124.43, 185.01)	**161.26*** (136.57, 198.55)	**184.62*** (155.03, 230.8)
P(Sleep-stage compactness)	%	76.97 (71.41, 78.93)	67.13 (53.56, 70.5)	64.53 (51.15, 67.77)	60.6 (47.5, 63.3)
	CATE		**-9.85*** (-19.52, -7.1)	**-12.44*** (-21.56, -9.78)	**-16.38*** (-24.97, -13.75)
	RR-CATE		**87.21*** (73.02, 90.71)	**83.84*** (70.12, 87.17)	**78.73*** (65.58, 82.08)
P(Sleep-stage fragmentation)	%	13.42 (11.04, 17.5)	19.16 (17.2, 29.1)	20.51 (18.28, 29.68)	22.55 (20.13, 31.66)
	CATE		**5.74*** (2.85, 14.85)	**7.09*** (3.97, 15.67)	**9.13*** (5.63, 17.11)
	RR-CATE		**142.78*** (120.55, 198.49)	**152.82*** (127.93, 205.73)	**168.03*** (139.01, 219.11)
P(W $\to$ W)	%	3.51 (1.82, 6.07)	4.8 (1.68, 7.56)	5.13 (1.75, 8.07)	5.6 (1.92, 8.91)
	CATE		1.29 (-1.47, 4.12)	1.62 (-1.28, 4.51)	2.09 (-0.93, 5.24)
	RR-CATE		136.65 (48.5, 264.53)	146.08 (53.57, 288.02)	159.46 (56.6, 309.79)
P(N1 $\to$ N1)	%	6.02 (4.46, 7.56)	5.2 (2.34, 6.61)	6.07 (2.68, 7.5)	7.55 (3.27, 9.26)
	CATE		-0.82 (-3.36, 0.8)	0.05 (-2.87, 1.72)	1.53 (-1.97, 3.47)
	RR-CATE		86.44 (43.49, 114.53)	100.81 (51.32, 131.2)	125.39 (65.66, 165.26)
P(N2 $\to$ N2)	%	37.55 (33.15, 41.94)	37.37 (32.08, 39.99)	35.73 (30.25, 37.95)	32.99 (27.23, 35.42)
	CATE		-0.18 (-7.32, 4.63)	-1.82 (-9.5, 3.01)	-4.55 (-12.92, 0.09)
	RR-CATE		99.52 (82.57, 113.95)	95.16 (75.88, 108.8)	87.87 (67.8, 100.27)
P(N3 $\to$ N3)	%	18.34 (7.43, 20.79)	14.86 (5.7, 17.68)	13.31 (4.96, 15.55)	11.15 (4.19, 13.2)
	CATE		-3.48 (-7.51, 0.55)	**-5.03*** (-9.05, -1.16)	**-7.19*** (-10.64, -2.35)
	RR-CATE		81.03 (49.63, 105.85)	**72.59*** (44.01, 91.8)	**60.77*** (39.8, 76.63)
P(REM $\to$ REM)	%	15.07 (13.69, 20.4)	9.7 (8.03, 14.05)	9.42 (7.99, 13.8)	8.91 (7.28, 13.01)
	CATE		**-5.37*** (-9.41, -2.3)	**-5.64*** (-9.83, -2.68)	**-6.15*** (-10.56, -3.05)
	RR-CATE		**64.36*** (50.52, 84.26)	**62.54*** (50.52, 81.8)	**59.15*** (45.33, 79.15)
P(W-fragmentation)	%	3.01 (2.62, 4)	4.2 (3.55, 6.48)	4.66 (4.02, 7.11)	5.39 (4.59, 8.52)
	CATE		**1.18*** (0.6, 2.7)	**1.65*** (0.99, 3.31)	**2.38*** (1.65, 4.44)
	RR-CATE		**139.27*** (119.02, 173.29)	**154.69*** (133.34, 190.31)	**178.86*** (152.82, 225.96)
P(N1-fragmentation)	%	5.45 (4.72, 6.8)	7.34 (6.58, 9.43)	8.18 (7.48, 10.29)	9.5 (8.7, 11.5)
	CATE		**1.89*** (0.88, 3.13)	**2.72*** (1.8, 3.98)	**4.05*** (3, 5.45)
	RR-CATE		**134.59*** (115.31, 159.43)	**149.96*** (131.13, 174.85)	**174.24*** (149.31, 198.76)
P(N2-fragmentation)	%	6.44 (5.32, 8.59)	8.97 (8, 14.33)	9.54 (8.51, 14.65)	10.41 (9.35, 15.37)
	CATE		**2.53*** (1.1, 6.61)	**3.1*** (1.66, 7.07)	**3.97*** (2.54, 7.94)
	RR-CATE		**139.3*** (116.03, 183.45)	**148.19*** (121.75, 191.06)	**161.68*** (135.07, 202.4)
P(N3-fragmentation)	%	2.93 (2.23, 4.29)	4.59 (3.88, 7.89)	4.73 (3.99, 8.03)	4.9 (4.15, 8.33)
	CATE		**1.66*** (0.59, 4.89)	**1.8*** (0.79, 4.81)	**1.97*** (0.93, 4.67)
	RR-CATE		**156.56*** (117.24, 230.22)	**161.32*** (122.44, 232.53)	**167.29*** (128.36, 238.25)
P(REM-fragmentation)	%	1.68 (1.19, 2.37)	2.98 (2.46, 4.7)	3.23 (2.72, 5.02)	3.6 (3, 5.37)
	CATE		**1.3*** (0.64, 3.01)	**1.55*** (0.84, 3.28)	**1.92*** (1.14, 3.67)
	RR-CATE		**177.4*** (131.47, 280.3)	**192.2*** (144.46, 301.09)	**214.16*** (157.16, 324.83)

Notes: Probabilities are expressed as percentages. Estimates include conditional average treatment effects (CATE) and risk-ratio CATE (RR-CATE).

Table B.4: Expected probabilities and estimated OSA effects (CATE, RR-CATE) for 70-year-old females.

Quantity	Estimate	Healthy	O1: OSA (AHI = 5)	O2: OSA (AHI = 15)	O3: OSA (AHI = 30)
P(W)	%	13.8 (7.73, 20.64)	18.74 (13.44, 22.39)	19.94 (14.68, 23.54)	21.63 (16.57, 24.99)
	CATE		4.94 (-4.07, 11.5)	6.14 (-2.46, 12.33)	7.83 (-0.68, 13.55)
	RR-CATE		135.8 (79.4, 225.94)	144.51 (87.08, 242.92)	156.75 (96.45, 267.43)
P(N1)	%	13.45 (10.83, 16.5)	13.68 (10.24, 16.48)	15.45 (11.91, 18.02)	18.33 (14.32, 20.73)
	CATE		0.23 (-3.09, 3.34)	2 (-1.78, 5.02)	**4.87*** (1.03, 8.02)
	RR-CATE		101.7 (76.75, 127.51)	114.84 (89.14, 140.83)	**136.23*** (105.89, 166.41)
P(N2)	%	41.67 (34.75, 46.54)	41.83 (37.94, 46.19)	40.36 (36.81, 44.91)	37.96 (34.94, 42.21)
	CATE		0.16 (-4.67, 7.61)	-1.3 (-6.17, 5.71)	-3.71 (-8.8, 4.44)
	RR-CATE		100.39 (89.51, 121.1)	96.87 (86.38, 116.94)	91.1 (80.64, 111.68)
P(N3)	%	18.69 (11.54, 34.87)	16.47 (11.19, 25.59)	15.06 (10.31, 24.49)	13.09 (9.34, 20.67)
	CATE		-2.21 (-11.12, 2.79)	-3.63 (-12.84, 0.61)	**-5.6*** (-15.5, -0.66)
	RR-CATE		88.16 (64.41, 115.1)	80.57 (61.66, 102.97)	**70.04*** (52.99, 93.94)
P(REM)	%	12.39 (6.86, 18.64)	9.27 (6.49, 13.55)	9.19 (6.58, 13.02)	8.99 (6.53, 12.45)
	CATE		-3.12 (-6.41, 0.65)	-3.21 (-6.65, 0.8)	-3.4 (-7.42, 0.65)
	RR-CATE		74.81 (61.93, 108.95)	74.13 (61.67, 110.74)	72.57 (58.98, 107.88)
P((N1,N2,N3,REM) $\rightarrow$ W)	%	2.89 (1.76, 4.04)	4.33 (3.19, 6.02)	4.72 (3.59, 6.39)	5.3 (4.15, 6.77)
	CATE		**1.44*** (0.81, 2.59)	**1.82*** (1.22, 2.96)	**2.4*** (1.81, 3.52)
	RR-CATE		**149.69*** (125.93, 205.77)	**162.94*** (136.15, 227.13)	**183.04*** (151.34, 257.14)
P((N1,N2) $\rightarrow$ W)	%	2.02 (1.25, 2.84)	3.05 (2.24, 4.08)	3.37 (2.56, 4.45)	3.87 (3.07, 4.82)
	CATE		**1.03*** (0.54, 1.84)	**1.35*** (0.92, 2.15)	**1.85*** (1.4, 2.64)
	RR-CATE		**151.07*** (122.94, 203.07)	**166.94*** (137.15, 228.97)	**191.67*** (155.42, 261.8)
P(N3 $\rightarrow$ W)	%	0.42 (0.25, 0.55)	0.57 (0.41, 0.74)	0.6 (0.45, 0.78)	0.64 (0.49, 0.81)
	CATE		**0.15*** (0.02, 0.35)	**0.18*** (0.05, 0.37)	**0.22*** (0.09, 0.41)
	RR-CATE		**134.54*** (103.84, 209.11)	**141.49*** (109.8, 218.61)	**150.68*** (116.73, 230.17)
P(REM $\rightarrow$ W)	%	0.45 (0.19, 0.75)	0.71 (0.47, 1.19)	0.75 (0.49, 1.17)	0.79 (0.54, 1.18)
	CATE		**0.26*** (0.06, 0.53)	**0.29*** (0.09, 0.56)	**0.34*** (0.12, 0.6)
	RR-CATE		**157.76*** (110.52, 334.5)	**165.22*** (115.34, 349.12)	**174.87*** (120.53, 385.77)
P(NREM $\rightleftarrows$ REM)	%	2.24 (1.27, 3.46)	3.82 (2.8, 5.02)	4.09 (3.05, 5.21)	4.49 (3.3, 5.38)
	CATE		**1.58*** (0.85, 2.52)	**1.86*** (1.02, 2.76)	**2.25*** (1.34, 3.09)
	RR-CATE		**170.81*** (129.14, 284.15)	**183.08*** (135.16, 303.51)	**200.68*** (143.42, 330.32)
P(N1 $\rightleftarrows$ N2)	%	5.76 (3.82, 7.61)	6.66 (5.39, 8.42)	7.39 (6.04, 9.22)	8.54 (7.14, 10.54)
	CATE		0.9 (-0.41, 2.98)	**1.63*** (0.37, 3.7)	**2.78*** (1.51, 4.79)
	RR-CATE		115.55 (94.17, 178)	**128.3*** (105.39, 196.79)	**148.27*** (122.41, 225.89)
P(Sleep compactness)	%	83.3 (76.47, 89.69)	77.47 (72.87, 83.36)	75.9 (71.72, 81.81)	73.65 (70.14, 79.12)
	CATE		-5.83 (-12.03, 3.03)	-7.4 (-13.69, 1.33)	**-9.65*** (-15.89, -0.76)
	RR-CATE		93 (86.19, 103.88)	91.12 (84.42, 101.74)	**88.42*** (81.88, 99.03)
P(Sleep fragmentation)	%	5.79 (3.58, 7.86)	8.12 (5.96, 11.21)	8.87 (6.72, 11.91)	10.01 (7.78, 12.72)
	CATE		**2.32*** (1.22, 4.35)	**3.08*** (2.01, 4.97)	**4.22*** (3.16, 6.22)
	RR-CATE		**140.11*** (118.84, 184.86)	**153.08*** (129.66, 204.25)	**172.79*** (144.49, 237.49)
P(Sleep-stage compactness)	%	71.21 (66.28, 79.03)	60.88 (54.79, 68.14)	58.35 (53.03, 65.4)	54.67 (50.01, 62.23)
	CATE		**-10.33*** (-15.96, -3.3)	**-12.87*** (-18.33, -6.4)	**-16.54*** (-21.92, -9.7)
	RR-CATE		**85.5*** (78.04, 94.81)	**81.93*** (75.07, 90.62)	**76.77*** (70.92, 86.32)
P(Sleep-stage fragmentation)	%	12.09 (7.7, 16.15)	16.59 (13.43, 21.14)	17.56 (14.4, 21.82)	18.98 (15.98, 22.68)
	CATE		**4.5*** (1.81, 9.37)	**5.47*** (2.72, 10.01)	**6.89*** (3.7, 11.3)
	RR-CATE		**137.23*** (112.42, 230.9)	**145.24*** (117.68, 234.03)	**157.03*** (126, 255.23)
P(W $\rightarrow$ W)	%	10.91 (4.79, 18.51)	14.41 (9.29, 17.2)	15.23 (10.04, 18.13)	16.33 (11.23, 19.13)
	CATE		3.5 (-5.93, 9.37)	4.32 (-5.09, 9.86)	5.43 (-3.51, 10.84)
	RR-CATE		132.12 (60.79, 257.6)	139.62 (67.22, 272.05)	149.77 (74.8, 292.61)
P(N1 $\rightarrow$ N1)	%	8.35 (6.57, 10.76)	6.98 (4.45, 8.43)	8.05 (5.22, 9.61)	9.83 (6.55, 11.54)
	CATE		-1.37 (-4.15, 0.72)	-0.31 (-3.22, 1.84)	1.48 (-1.84, 3.87)
	RR-CATE		83.57 (49.73, 110.6)	96.35 (60.43, 125.95)	117.76 (79.01, 155.24)
P(N2 $\rightarrow$ N2)	%	35.85 (29.77, 40.35)	34.5 (30.6, 38.71)	32.61 (29.19, 36.5)	29.59 (26.58, 33.06)
	CATE		-1.36 (-6.04, 4.57)	-3.25 (-7.91, 2.61)	-6.26 (-11.26, 0.82)
	RR-CATE		96.22 (83.82, 114.11)	90.95 (79.1, 108.59)	82.53 (70.26, 102.74)
P(N3 $\rightarrow$ N3)	%	16.14 (9.49, 33.3)	12.64 (6.87, 22.38)	11.2 (6.08, 21.43)	9.21 (5.23, 17.07)
	CATE		-3.5 (-12.45, 1.11)	**-4.94*** (-13.98, -1.32)	**-6.93*** (-16.72, -2.62)
	RR-CATE		78.34 (56.09, 106.72)	**69.37*** (50.76, 91.53)	**57.08*** (42.04, 78.53)
P(REM $\rightarrow$ REM)	%	10.87 (5.9, 17.05)	6.76 (4.53, 10.91)	6.49 (4.37, 10.14)	6.04 (4.24, 9.38)
	CATE		**-4.1*** (-7.73, -0.55)	**-4.37*** (-8.19, -0.58)	**-4.83*** (-8.96, -0.85)
	RR-CATE		**62.23*** (49.83, 91.26)	**59.77*** (48.03, 93.54)	**55.55*** (42.97, 87.33)
P(W-fragmentation)	%	2.9 (1.82, 3.94)	3.79 (2.77, 5.17)	4.15 (3.13, 5.4)	4.71 (3.68, 5.94)
	CATE		**0.89*** (0.36, 1.78)	**1.25*** (0.75, 2.08)	**1.81*** (1.34, 2.74)
	RR-CATE		**130.55*** (110.34, 171.89)	**143.23*** (122.03, 184.49)	**162.55*** (138.65, 218.91)
P(N1-fragmentation)	%	5.39 (3.84, 6.86)	7.08 (5.69, 8.61)	7.8 (6.51, 9.36)	8.92 (7.62, 10.35)
	CATE		**1.69*** (0.72, 3.21)	**2.41*** (1.45, 3.95)	**3.53*** (2.56, 5.23)
	RR-CATE		**131.36*** (111.36, 185.69)	**144.83*** (122.94, 205.3)	**165.62*** (139.92, 231.63)
P(N2-fragmentation)	%	5.68 (3.42, 7.75)	7.64 (6.11, 10)	8.04 (6.58, 10.24)	8.64 (7.19, 10.72)
	CATE		**1.96*** (0.71, 4.31)	**2.36*** (1.12, 4.55)	**2.96*** (1.69, 5.05)
	RR-CATE		**134.4*** (111.99, 229.52)	**141.51*** (114.46, 238.97)	**152.04*** (124.4, 253)
P(N3-fragmentation)	%	2.54 (1.52, 3.79)	3.83 (3.04, 5.18)	3.88 (3.08, 5.09)	3.93 (3.13, 4.93)
	CATE		**1.29*** (0.3, 2.71)	**1.34*** (0.37, 2.65)	**1.39*** (0.45, 2.55)
	RR-CATE		**150.72*** (109.06, 278.93)	**152.91*** (110.17, 275.26)	**154.84*** (112.07, 268.2)
P(REM-fragmentation)	%	1.37 (0.74, 2.21)	2.38 (1.73, 3.27)	2.55 (1.9, 3.38)	2.79 (2.02, 3.53)
	CATE		**1.01*** (0.42, 1.62)	**1.17*** (0.55, 1.78)	**1.41*** (0.74, 2.01)
	RR-CATE		**173.23*** (123.6, 301.1)	**185.4*** (131.69, 314.71)	**202.79*** (142.65, 332.9)

Notes: Probabilities are expressed as percentages. Estimates include conditional average treatment effects (CATE) and risk-ratio CATE (RR-CATE).

Table B.5: Expected probabilities and estimated OSA effects (CATE, RR-CATE) for 30-year-old males.

Quantity	Estimate	Healthy	O1: OSA (AHI = 5)	O2: OSA (AHI = 15)	O3: OSA (AHI = 30)
P(W)	%	6.24 (4.41, 9.27)	6.6 (4.59, 11.36)	7.16 (4.95, 11.89)	8.01 (5.61, 12.25)
	CATE		0.35 (-1.68, 4.55)	0.92 (-1.05, 5.07)	1.76 (-0.36, 6.27)
	RR-CATE		105.69 (76.98, 186.83)	114.68 (85.13, 201.2)	128.26 (94.34, 224.75)
P(N1)	%	11.51 (8.83, 14)	12.42 (10.39, 14.18)	14.09 (11.59, 15.7)	16.85 (13.59, 18.63)
	CATE		0.9 (-0.97, 3.82)	**2.57*** (0.66, 5.26)	**5.33*** (2.66, 8.02)
	RR-CATE		107.85 (91.92, 140.3)	**122.36*** (105.28, 159.89)	**146.31*** (120.47, 187.6)
P(N2)	%	43.53 (35.5, 48.16)	47.15 (31.96, 49.36)	46.11 (31.57, 48.49)	44.3 (31.25, 47.09)
	CATE		3.62 (-4.01, 6.34)	2.57 (-4.99, 5.7)	0.77 (-7.05, 3.88)
	RR-CATE		108.3 (90.39, 116.23)	105.91 (87.91, 113.33)	101.76 (84.46, 109.71)
P(N3)	%	16.93 (5.67, 21.61)	16.1 (9.13, 23.13)	15.03 (8.94, 21.79)	13.53 (8.7, 19.75)
	CATE		-0.83 (-5.03, 7.79)	-1.91 (-6.21, 6.34)	-3.41 (-7.25, 5.82)
	RR-CATE		95.1 (72.79, 173.01)	88.74 (68.08, 173.22)	79.88 (62.77, 178.25)
P(REM)	%	21.77 (17.6, 38.26)	17.73 (12.45, 32.81)	17.62 (12.51, 32.06)	17.32 (12.07, 31.15)
	CATE		-4.04 (-10.82, 1)	-4.16 (-10.92, 0.54)	**-4.46*** (-11.44, -0.25)
	RR-CATE		81.43 (57.64, 105.31)	80.91 (57.82, 102.88)	**79.54*** (54.34, 98.67)
P((N1,N2,N3,REM) $\to$ W)	%	3.97 (2.74, 5.53)	4.94 (3.27, 9.76)	5.39 (3.57, 10.25)	6.08 (4.03, 10.95)
	CATE		0.97 (-0.16, 4.48)	**1.42*** (0.2, 5)	**2.11*** (0.66, 5.33)
	RR-CATE		124.42 (95.66, 197.26)	**135.75*** (105.9, 209.74)	**153.11*** (119.35, 219.6)
P((N1,N2) $\to$ W)	%	2.65 (2.01, 4)	3.28 (2.17, 7.33)	3.63 (2.39, 7.66)	4.18 (2.78, 7.96)
	CATE		0.64 (-0.14, 3.1)	**0.99*** (0.11, 3.55)	**1.54*** (0.47, 3.94)
	RR-CATE		123.99 (94.43, 189.08)	**137.23*** (105.04, 200)	**158.07*** (120.57, 219.84)
P(N3 $\to$ W)	%	0.57 (0.3, 0.88)	0.72 (0.52, 1.25)	0.76 (0.56, 1.26)	0.83 (0.59, 1.31)
	CATE		0.14 (-0.04, 0.62)	**0.19*** (0.01, 0.67)	**0.25*** (0.1, 0.7)
	RR-CATE		125.09 (94.73, 248.93)	**132.88*** (100.93, 256.6)	**143.77*** (115.55, 267.93)
P(REM $\to$ W)	%	0.75 (0.36, 1.32)	0.94 (0.49, 2.19)	0.99 (0.53, 2.21)	1.07 (0.55, 2.04)
	CATE		0.19 (-0.26, 1.06)	0.24 (-0.21, 1.06)	0.32 (-0.14, 1.06)
	RR-CATE		125.45 (68.89, 278.93)	132.72 (72.47, 279.01)	142.71 (75.83, 275.8)
P(NREM $\rightleftarrows$ REM)	%	4.19 (2.29, 6)	5.75 (4.26, 14.95)	6.22 (4.57, 15.85)	6.91 (5.18, 17.1)
	CATE		**1.56*** (0.63, 10.4)	**2.03*** (1.08, 11.05)	**2.73*** (1.64, 11.75)
	RR-CATE		**137.26*** (115.38, 329.3)	**148.46*** (126.77, 337.19)	**165.11*** (138.22, 348.28)
P(N1 $\rightleftarrows$ N2)	%	6.74 (5.51, 9.8)	8.04 (6.05, 10.06)	9.02 (6.83, 10.91)	10.6 (8.1, 12.59)
	CATE		1.3 (-0.93, 2.28)	2.28 (-0.41, 3.28)	**3.86*** (0.27, 4.94)
	RR-CATE		119.35 (88.57, 136.32)	133.89 (95.73, 153.19)	**157.27*** (102.64, 182.3)
P(Sleep compactness)	%	90.09 (86.76, 93.12)	88.91 (77.31, 92.11)	87.86 (75.78, 91.49)	86.23 (73.58, 90.35)
	CATE		-1.18 (-12.26, 1.46)	-2.23 (-13.14, 0.35)	**-3.85*** (-14.94, -0.97)
	RR-CATE		98.69 (86.34, 101.66)	97.52 (85.03, 100.39)	**95.72*** (83.55, 98.9)
P(Sleep fragmentation)	%	7.64 (5.14, 10.41)	9.43 (6.36, 22.33)	10.37 (7, 23.46)	11.83 (7.99, 25.47)
	CATE		1.79 (-0.15, 12.7)	**2.73*** (0.69, 14.03)	**4.2*** (1.85, 15.68)
	RR-CATE		123.47 (97.84, 231.34)	**135.8*** (109.56, 244.75)	**154.96*** (126.99, 267.12)
P(Sleep-stage compactness)	%	74.39 (65.36, 80.53)	68.28 (39.47, 74.61)	65.7 (37.23, 71.94)	61.74 (33.12, 68.45)
	CATE		**-6.11*** (-28.22, -1.72)	**-8.7*** (-31.06, -4.67)	**-12.66*** (-34.15, -8.66)
	RR-CATE		**91.78*** (57.61, 97.68)	**88.31*** (53.79, 93.69)	**82.99*** (49.69, 88.56)
P(Sleep-stage fragmentation)	%	15.7 (12.01, 22.29)	20.63 (16.3, 37.53)	22.16 (17.53, 38.56)	24.5 (19.28, 40.46)
	CATE		**4.93*** (2.07, 15.87)	**6.46*** (3.51, 17.25)	**8.8*** (5.62, 18.57)
	RR-CATE		**131.44*** (112.95, 178.98)	**141.18*** (122.93, 180.9)	**156.07*** (137.57, 196.56)
P(W $\to$ W)	%	2.27 (0.56, 4.3)	1.66 (0.2, 5.04)	1.77 (0.21, 5.48)	1.93 (0.21, 5.95)
	CATE		-0.61 (-2.04, 2.34)	-0.5 (-1.96, 2.83)	-0.34 (-1.86, 3.53)
	RR-CATE		72.97 (27.74, 209.34)	77.9 (28.66, 222.01)	84.89 (30.26, 252.4)
P(N1 $\to$ N1)	%	5.17 (2.5, 7.02)	4.7 (1.2, 6.01)	5.47 (1.29, 7.09)	6.8 (1.45, 8.75)
	CATE		-0.47 (-2.46, 1.46)	0.31 (-2.03, 2.41)	1.63 (-1.27, 4.05)
	RR-CATE		90.98 (45.72, 136.63)	105.94 (52.65, 156.15)	131.56 (67.16, 194.25)
P(N2 $\to$ N2)	%	36.43 (25.04, 41.75)	38.02 (14.13, 41.41)	36.3 (13.4, 39.8)	33.47 (10.67, 37.31)
	CATE		1.59 (-11.16, 4.74)	-0.13 (-12.11, 2.87)	**-2.96*** (-13.94, -0.23)
	RR-CATE		104.37 (56.45, 114.73)	99.65 (54.01, 107.84)	**91.87*** (45.9, 99.37)
P(N3 $\to$ N3)	%	13.74 (1.39, 19.1)	11.52 (0.29, 19.09)	10.3 (0.27, 17.7)	8.61 (0.24, 15.4)
	CATE		-2.23 (-8.07, 6.21)	-3.44 (-8.55, 4.26)	-5.13 (-10.17, 1.88)
	RR-CATE		83.81 (18.41, 155.79)	74.96 (16.25, 139.74)	62.66 (13.37, 116.7)
P(REM $\to$ REM)	%	19.05 (15.74, 34.09)	14.04 (9.12, 23.19)	13.62 (8.99, 22.08)	12.86 (8.21, 20.54)
	CATE		**-5.01*** (-12.62, -0.82)	**-5.43*** (-13.65, -1.78)	**-6.19*** (-15.34, -2.91)
	RR-CATE		**73.68*** (46.42, 95.58)	**71.48*** (44.74, 89.62)	**67.5*** (41.94, 83.34)
P(W-fragmentation)	%	3.67 (2.44, 4.71)	4.49 (3.07, 12.63)	4.98 (3.43, 13.44)	5.76 (3.98, 14.9)
	CATE		0.82 (-0.09, 8.2)	**1.32*** (0.43, 9.21)	**2.09*** (1.08, 10.29)
	RR-CATE		122.43 (97, 282.77)	**135.86*** (111.56, 299.2)	**156.97*** (132.91, 323.74)
P(N1-fragmentation)	%	6.46 (5.06, 8.57)	7.64 (6.16, 10.09)	8.5 (6.91, 10.92)	9.85 (8.09, 12.24)
	CATE		**1.18*** (0.26, 2.35)	**2.04*** (1.07, 3.08)	**3.39*** (2.28, 4.4)
	RR-CATE		**118.29*** (104.19, 140.53)	**131.5*** (117.26, 154.57)	**152.39*** (133.8, 174.95)
P(N2-fragmentation)	%	7.46 (5.96, 11.83)	9.76 (7.27, 16.94)	10.43 (7.77, 17.27)	11.45 (8.41, 17.67)
	CATE		**2.3*** (0.54, 5.29)	**2.97*** (0.94, 5.77)	**4*** (1.55, 6.24)
	RR-CATE		**130.91*** (107.81, 162.45)	**139.87*** (113.93, 165.49)	**153.61*** (122.92, 181.7)
P(N3-fragmentation)	%	3.13 (2.19, 4.81)	4.58 (3.38, 10.81)	4.74 (3.51, 10.99)	4.96 (3.7, 11.26)
	CATE		**1.45*** (0.69, 6.26)	**1.61*** (0.86, 6.41)	**1.82*** (1.11, 6.42)
	RR-CATE		**146.22*** (121.73, 238.07)	**151.37*** (128.16, 236.6)	**158.23*** (136.63, 243.25)
P(REM-fragmentation)	%	2.62 (1.31, 3.75)	3.59 (2.54, 10.36)	3.88 (2.75, 10.83)	4.32 (3.06, 11.6)
	CATE		**0.97*** (0.16, 7.21)	**1.26*** (0.48, 7.82)	**1.7*** (0.87, 8.38)
	RR-CATE		**137.07*** (105.93, 322.35)	**148.33*** (117.64, 343.87)	**165.09*** (133.09, 368.89)

Notes: Probabilities are expressed as percentages. Estimates include conditional average treatment effects (CATE) and risk-ratio CATE (RR-CATE).

Table B.6: Expected probabilities and estimated OSA effects (CATE, RR-CATE) for 50-year-old females.

Quantity	Estimate	Healthy	O1: OSA (AHI = 5)	O2: OSA (AHI = 15)	O3: OSA (AHI = 30)
P(W)	%	11.24 (6.32, 14.97)	10.23 (8.07, 12.87)	10.99 (8.65, 13.52)	12.08 (9.58, 14.94)
	CATE		-1 (-4.77, 5.2)	-0.25 (-4.31, 6.31)	0.84 (-3.04, 7.9)
	RR-CATE		91.08 (67.3, 179.25)	97.77 (70.51, 196.92)	107.52 (78.54, 216.37)
P(N1)	%	13.92 (11.14, 16.57)	14.65 (12.98, 16.52)	16.57 (14.99, 18.22)	19.71 (18.14, 21.23)
	CATE		0.73 (-1.52, 3.8)	**2.65*** (0.5, 5.49)	**5.79*** (3.49, 8.58)
	RR-CATE		105.21 (90.75, 134.95)	**119.02*** (103.32, 150.89)	**141.56*** (123.49, 176.82)
P(N2)	%	42.78 (40.29, 47.62)	46.56 (42.59, 48.67)	45.15 (41.37, 46.85)	42.79 (38.66, 44.68)
	CATE		3.79 (-3.26, 6.64)	2.37 (-4.88, 4.75)	0.01 (-7.33, 1.96)
	RR-CATE		108.85 (93.17, 116.39)	105.55 (89.76, 111.91)	100.03 (85.14, 104.85)
P(N3)	%	15.41 (8.24, 19.16)	14.72 (7.67, 17.32)	13.61 (7.51, 15.62)	12.06 (7.73, 13.75)
	CATE		-0.69 (-6.21, 6.17)	-1.81 (-6.96, 4.71)	-3.36 (-8.05, 2.76)
	RR-CATE		95.51 (65.63, 161.66)	88.29 (63, 147.19)	78.23 (57.57, 126.81)
P(REM)	%	16.65 (13.92, 23)	13.83 (12.06, 21.33)	13.69 (12.13, 21.48)	13.37 (11.58, 20.8)
	CATE		-2.82 (-9.2, 2.94)	-2.96 (-9.35, 2.55)	-3.29 (-9.84, 2.05)
	RR-CATE		83.08 (58.65, 116.08)	82.2 (58.03, 113.47)	80.26 (55.27, 112.83)
P((N1,N2,N3,REM) $\rightarrow$ W)	%	3.89 (2.81, 4.89)	4.83 (4.08, 7.49)	5.25 (4.52, 8.01)	5.9 (5.1, 8.7)
	CATE		0.93 (-0.18, 4.31)	**1.36*** (0.22, 4.7)	**2.01*** (0.73, 5.36)
	RR-CATE		123.93 (95.69, 233.47)	**134.9*** (104.54, 246.1)	**151.53*** (116.55, 269.71)
P((N1,N2) $\rightarrow$ W)	%	2.72 (2.12, 3.28)	3.34 (2.81, 5.1)	3.68 (3.17, 5.59)	4.23 (3.66, 6.27)
	CATE		0.62 (-0.12, 2.8)	**0.97*** (0.22, 3.12)	**1.51*** (0.68, 3.77)
	RR-CATE		122.82 (96.02, 214.24)	**135.67*** (107.12, 226.57)	**155.7*** (121.47, 257.59)
P(N3 $\rightarrow$ W)	%	0.49 (0.27, 0.6)	0.62 (0.5, 0.91)	0.65 (0.53, 0.92)	0.7 (0.59, 0.95)
	CATE		0.13 (-0.04, 0.56)	0.16 (0, 0.57)	**0.21*** (0.08, 0.59)
	RR-CATE		126.26 (94.09, 285.85)	**133.21*** (100.83, 293.09)	**142.55*** (113.48, 289.69)
P(REM $\rightarrow$ W)	%	0.69 (0.38, 1.04)	0.87 (0.69, 1.6)	0.92 (0.74, 1.64)	0.98 (0.78, 1.57)
	CATE		0.18 (-0.29, 0.99)	0.23 (-0.26, 1.05)	0.29 (-0.25, 1.1)
	RR-CATE		126.63 (69.68, 327.29)	133.05 (74.48, 338.65)	141.5 (76.35, 344.74)
P(NREM $\rightleftarrows$ REM)	%	3.57 (2.06, 4.17)	4.92 (4.17, 7.93)	5.29 (4.63, 8.25)	5.82 (5.13, 9.05)
	CATE		**1.36*** (0.59, 5.02)	**1.72*** (1.03, 5.45)	**2.25*** (1.39, 6.12)
	RR-CATE		**138.03*** (115.19, 322.55)	**148.3*** (127.19, 333.03)	**163.19*** (135.24, 347.69)
P(N1 $\rightleftarrows$ N2)	%	6.67 (5.58, 8.53)	8.03 (7.22, 11.05)	8.95 (8.15, 11.9)	10.4 (9.44, 13.2)
	CATE		1.37 (-0.64, 3.14)	**2.28*** (0.21, 4.01)	**3.73*** (1.59, 5.37)
	RR-CATE		120.48 (92.28, 145.73)	**134.24*** (102.62, 160.17)	**155.96*** (119.33, 184.75)
P(Sleep compactness)	%	85.13 (81.26, 90.66)	85.49 (81.46, 87.39)	84.31 (80.57, 86.34)	82.55 (79.16, 84.74)
	CATE		0.36 (-6.99, 4.58)	-0.83 (-8.26, 3.63)	-2.59 (-10.21, 2.07)
	RR-CATE		100.42 (92.38, 105.67)	99.03 (90.93, 104.45)	96.96 (88.7, 102.55)
P(Sleep fragmentation)	%	7.52 (5.6, 9.16)	9.1 (7.75, 14.61)	9.96 (8.62, 15.84)	11.27 (9.77, 17.24)
	CATE		1.58 (-0.46, 8.5)	**2.44*** (0.41, 9.44)	**3.75*** (1.43, 10.52)
	RR-CATE		120.95 (94.71, 231.03)	**132.38*** (104.62, 244.07)	**149.83*** (116.84, 268.84)
P(Sleep-stage compactness)	%	70.53 (66.74, 76.39)	66.25 (53.63, 69.03)	63.77 (51.35, 66.39)	60.08 (47.63, 62.8)
	CATE		-4.28 (-18.11, 0.4)	**-6.75*** (-19.95, -2.64)	**-10.45*** (-22.81, -6.6)
	RR-CATE		93.94 (75.3, 100.6)	**90.42*** (71.73, 96.17)	**85.19*** (67.67, 90.29)
P(Sleep-stage fragmentation)	%	14.6 (12.21, 16.73)	19.24 (17.38, 28.47)	20.53 (18.59, 29.25)	22.46 (20.37, 30.79)
	CATE		**4.63*** (1.93, 12.96)	**5.93*** (3.34, 14)	**7.86*** (5.18, 15.38)
	RR-CATE		**131.74*** (112.14, 182.36)	**140.58*** (120.97, 188.27)	**153.82*** (132.46, 200.52)
P(W $\rightarrow$ W)	%	7.34 (3.15, 10.9)	5.41 (1.99, 7.71)	5.73 (2.1, 8.4)	6.18 (2.33, 9.09)
	CATE		-1.93 (-5.73, 3.84)	-1.61 (-5.5, 4.5)	-1.16 (-5.04, 5.34)
	RR-CATE		73.66 (36.6, 213.54)	78.09 (38.66, 223.98)	84.17 (41.58, 239.56)
P(N1 $\rightarrow$ N1)	%	7.46 (5.36, 9.7)	6.85 (4.4, 7.99)	7.92 (5.15, 8.92)	9.73 (5.87, 10.89)
	CATE		-0.61 (-3.63, 2)	0.46 (-2.42, 2.93)	2.27 (-0.9, 4.96)
	RR-CATE		91.83 (60.77, 136.2)	106.21 (71.69, 153.63)	130.45 (87.87, 191.07)
P(N2 $\rightarrow$ N2)	%	36.19 (33.54, 40.89)	38.13 (29.83, 39.94)	36.16 (27.67, 37.6)	32.97 (24.66, 34.48)
	CATE		1.94 (-7.08, 4.81)	-0.04 (-8.8, 2.31)	**-3.22*** (-12.03, -1)
	RR-CATE		105.35 (80.47, 114.4)	99.9 (76.64, 106.67)	**91.09*** (68.65, 97.12)
P(N3 $\rightarrow$ N3)	%	12.58 (5.16, 16.85)	10.64 (1.77, 13.24)	9.46 (1.68, 11.51)	7.82 (1.25, 9.85)
	CATE		-1.94 (-8.96, 4.8)	-3.13 (-9.45, 3.27)	-4.77 (-10.57, 1.2)
	RR-CATE		84.59 (26.11, 160.79)	75.15 (23.56, 142.39)	62.13 (20, 116.59)
P(REM $\rightarrow$ REM)	%	14.29 (12.02, 20.89)	10.63 (9.07, 15.82)	10.24 (8.83, 16.01)	9.57 (7.98, 14.74)
	CATE		-3.66 (-10.28, 0.27)	**-4.05*** (-10.69, -0.47)	**-4.73*** (-11.41, -1.35)
	RR-CATE		74.37 (47.68, 102.02)	**71.66*** (45.58, 97.18)	**66.93*** (42.1, 91.3)
P(W-fragmentation)	%	3.63 (2.76, 4.24)	4.27 (3.6, 7.24)	4.71 (4.01, 7.9)	5.37 (4.65, 8.81)
	CATE		0.64 (-0.31, 3.97)	**1.08*** (0.12, 4.68)	**1.74*** (0.71, 5.23)
	RR-CATE		117.75 (91.99, 223.8)	**129.68*** (102.91, 242.99)	**148.01*** (118.69, 265.48)
P(N1-fragmentation)	%	6.68 (5.68, 7.58)	7.92 (7.11, 10.71)	8.75 (7.98, 11.37)	10.05 (9.2, 12.87)
	CATE		**1.24*** (0.19, 3.22)	**2.08*** (1.1, 3.96)	**3.37*** (2.28, 5.24)
	RR-CATE		**118.65*** (102.67, 150.56)	**131.12*** (114.5, 163.56)	**150.49*** (132.49, 183.02)
P(N2-fragmentation)	%	6.81 (5.74, 8.11)	8.98 (8.04, 13.65)	9.54 (8.61, 14.16)	10.39 (9.41, 14.75)
	CATE		**2.17*** (0.54, 6.23)	**2.74*** (0.96, 6.73)	**3.58*** (1.51, 7.4)
	RR-CATE		**131.94*** (106.68, 180.12)	**140.2*** (112.16, 187.98)	**152.65*** (119.04, 199.3)
P(N3-fragmentation)	%	2.8 (2.07, 3.14)	4.1 (3.53, 6.75)	4.19 (3.65, 6.75)	4.31 (3.79, 6.97)
	CATE		**1.3*** (0.66, 4.02)	**1.4*** (0.74, 4.15)	**1.51*** (0.85, 4.05)
	RR-CATE		**146.53*** (122.09, 247.27)	**149.99*** (126, 245.88)	**154*** (129.54, 235.46)
P(REM-fragmentation)	%	2.22 (1.12, 2.71)	3.07 (2.62, 5.16)	3.29 (2.82, 5.3)	3.62 (3.13, 5.72)
	CATE		**0.85*** (0.17, 3.46)	**1.07*** (0.4, 3.63)	**1.4*** (0.72, 4.01)
	RR-CATE		**138.16*** (107.36, 329.54)	**148.39*** (116.67, 340.06)	**163.17*** (128.56, 342.75)

Notes: Probabilities are expressed as percentages. Estimates include conditional average treatment effects (CATE) and risk-ratio CATE (RR-CATE).

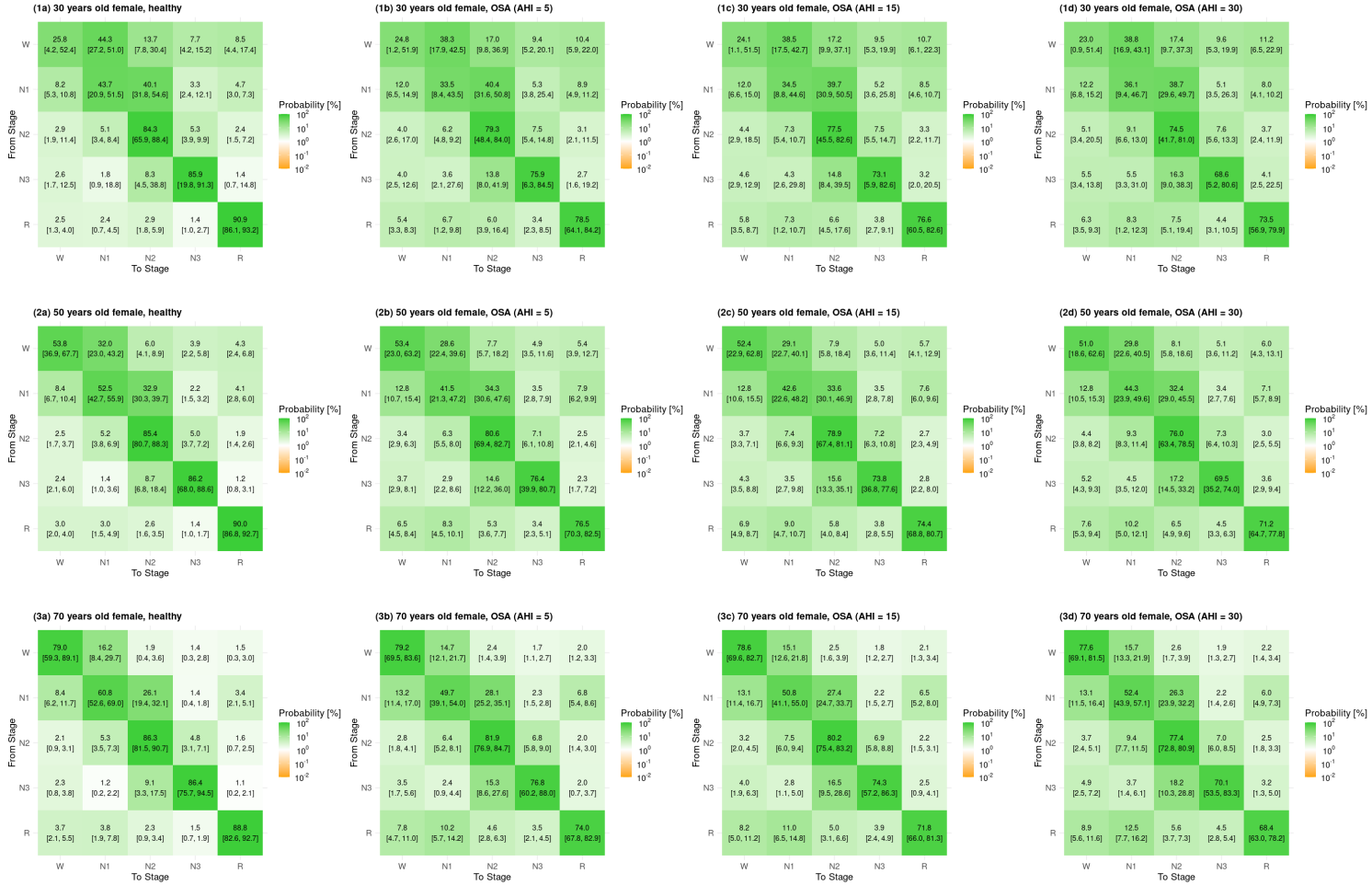
Table B.7: Expected probabilities and estimated OSA effects (CATE, RR-CATE) for 70-year-old females.

Quantity	Estimate	Healthy	O1: OSA (AHI = 5)	O2: OSA (AHI = 15)	O3: OSA (AHI = 30)
P(W)	%	24.25 (9.24, 31.04)	20.28 (14.57, 23.61)	21.39 (15.68, 24.18)	22.89 (17.57, 25.22)
	CATE		-3.98 (-12.28, 7.19)	-2.87 (-10.6, 8.48)	-1.37 (-8.93, 10.23)
	RR-CATE		83.61 (58.88, 179.93)	88.17 (64.35, 190.48)	94.36 (70.98, 208.42)
P(N1)	%	15.26 (12.59, 17.75)	16.17 (14.35, 19.68)	18.14 (16.67, 21.86)	21.3 (19.9, 25.12)
	CATE		0.91 (-2.07, 4.87)	2.88 (0, 7.05)	**6.04*** (3.33, 10.05)
	RR-CATE		105.97 (88.64, 135.5)	**118.86*** (100.01, 148.98)	**139.58*** (118.92, 173.61)
P(N2)	%	36.95 (34.21, 45.05)	41.6 (38.29, 45.66)	39.8 (37.04, 43.64)	36.96 (34.69, 40.61)
	CATE		4.65 (-2.41, 10.6)	2.85 (-3.78, 8.58)	0.01 (-6.45, 4.89)
	RR-CATE		112.58 (94.63, 130.46)	107.71 (91.72, 124.14)	100.03 (85.57, 114.11)
P(N3)	%	12.32 (6.17, 19.21)	12.15 (8.56, 15.05)	11.07 (7.99, 13.4)	9.6 (7.31, 11.37)
	CATE		-0.16 (-9.43, 4.58)	-1.24 (-9.87, 3.39)	-2.72 (-10.59, 1.97)
	RR-CATE		98.69 (49.31, 162.84)	89.92 (46.76, 147.29)	77.91 (42.11, 130)
P(REM)	%	11.22 (6.3, 21.78)	9.8 (8.31, 13.78)	9.61 (8.34, 13.08)	9.26 (8.03, 12.03)
	CATE		-1.42 (-8.22, 3.48)	-1.61 (-8.95, 3.42)	-1.96 (-10.27, 3.21)
	RR-CATE		87.32 (58.31, 155.78)	85.61 (57.35, 154.62)	82.5 (53.48, 150.53)
P((N1,N2,N3,REM) $\to$ W)	%	3.46 (1.76, 5.31)	4.39 (3.53, 5.99)	4.75 (3.96, 6.23)	5.28 (4.5, 6.55)
	CATE		0.93 (-0.1, 3.15)	**1.28*** (0.3, 3.38)	**1.81*** (0.92, 3.83)
	RR-CATE		126.78 (97.65, 274.51)	**137.01*** (107.01, 287.19)	**152.28*** (118.33, 318.1)
P((N1,N2) $\to$ W)	%	2.54 (1.44, 3.7)	3.18 (2.54, 4.22)	3.49 (2.92, 4.45)	3.96 (3.46, 4.87)
	CATE		**0.64*** (0.01, 1.93)	**0.95*** (0.32, 2.27)	**1.42*** (0.8, 2.81)
	RR-CATE		**125.16*** (100.25, 234.19)	**137.22*** (109.17, 259.82)	**155.72*** (122.68, 293.63)
P(N3 $\to$ W)	%	0.36 (0.14, 0.46)	0.48 (0.37, 0.59)	0.5 (0.39, 0.6)	0.52 (0.42, 0.62)
	CATE		0.11 (-0.01, 0.37)	**0.13*** (0.02, 0.41)	**0.16*** (0.06, 0.42)
	RR-CATE		130.99 (97.6, 352.35)	**136.52*** (105.43, 363.08)	**143.46*** (112.55, 385.68)
P(REM $\to$ W)	%	0.56 (0.17, 1.17)	0.74 (0.55, 1.17)	0.76 (0.58, 1.16)	0.8 (0.61, 1.14)
	CATE		0.18 (-0.24, 0.55)	0.2 (-0.23, 0.57)	0.24 (-0.23, 0.6)
	RR-CATE		131.38 (77.78, 452.58)	136.36 (78.9, 473.86)	142.4 (80.22, 447.72)
P(NREM $\rightleftarrows$ REM)	%	2.72 (1.42, 3.48)	3.88 (3.03, 4.74)	4.12 (3.35, 4.9)	4.45 (3.82, 5.22)
	CATE		**1.16*** (0.55, 2.75)	**1.4*** (0.82, 2.86)	**1.73*** (1.09, 3.12)
	RR-CATE		**142.56*** (117.62, 277.57)	**151.35*** (125.04, 290.05)	**163.59*** (133.12, 320.97)
P(N1 $\rightleftarrows$ N2)	%	5.78 (4.38, 9.42)	7.23 (6.49, 9.17)	7.96 (7.23, 9.96)	9.08 (8.4, 11.09)
	CATE		1.45 (-0.8, 3.93)	2.17 (-0.02, 4.56)	**3.3*** (1.34, 5.53)
	RR-CATE		125.01 (91.53, 189.09)	137.6 (99.81, 201.05)	**156.98*** (113.74, 223.75)
P(Sleep compactness)	%	72.49 (64.61, 87.34)	75.9 (71.91, 81.17)	74.46 (71.08, 79.54)	72.46 (69.58, 77.56)
	CATE		3.41 (-7.93, 12.57)	1.97 (-9.56, 10.49)	-0.03 (-11.87, 8.2)
	RR-CATE		104.7 (90.85, 118.97)	102.72 (88.99, 115.82)	99.96 (86.31, 112.71)
P(Sleep fragmentation)	%	6.72 (3.57, 10.33)	8.21 (6.6, 10.69)	8.9 (7.43, 11.21)	9.92 (8.49, 12.25)
	CATE		1.49 (-0.42, 5.25)	**2.18*** (0.37, 5.76)	**3.21*** (1.34, 6.81)
	RR-CATE		122.22 (95.17, 250.28)	**132.44*** (104.15, 261.75)	**147.7*** (113.64, 289.47)
P(Sleep-stage compactness)	%	60.46 (53.83, 71.54)	59.56 (55, 64.01)	57.23 (53.36, 61.45)	53.93 (50.72, 58.13)
	CATE		-0.9 (-11.58, 5.88)	-3.23 (-13.72, 2.9)	-6.53 (-16.86, 0.31)
	RR-CATE		98.51 (83.6, 110.46)	94.65 (80.89, 105.56)	89.2 (75.93, 100.6)
P(Sleep-stage fragmentation)	%	12.03 (8.14, 17.09)	16.34 (14.29, 19.79)	17.24 (15.23, 20.49)	18.53 (17.02, 21.52)
	CATE		**4.31*** (1.79, 10.1)	**5.2*** (2.77, 10.34)	**6.5*** (3.98, 10.88)
	RR-CATE		**135.82*** (111.39, 224.05)	**143.25*** (116.4, 226.95)	**154.04*** (124.69, 239.48)
P(W $\to$ W)	%	20.79 (5.55, 26.66)	15.89 (10.1, 18.62)	16.64 (10.91, 19.14)	17.61 (12.07, 19.78)
	CATE		-4.9 (-13.62, 5.95)	-4.15 (-12.72, 6.81)	-3.18 (-10.73, 7.89)
	RR-CATE		76.42 (50.56, 209.13)	80.03 (53.06, 222.71)	84.71 (57.23, 234.23)
P(N1 $\to$ N1)	%	9.43 (6.19, 11.12)	8.99 (7.34, 11.23)	10.27 (8.68, 12.48)	12.39 (10.81, 15.18)
	CATE		-0.45 (-3.15, 2.45)	0.84 (-1.6, 3.6)	**2.95*** (0.48, 5.59)
	RR-CATE		95.27 (72.01, 136.91)	108.85 (86.65, 154.44)	**131.28*** (104.19, 189.99)
P(N2 $\to$ N2)	%	31.52 (28.89, 37.56)	34.45 (30.83, 38.17)	32.27 (29.1, 35.36)	28.89 (26.32, 31.94)
	CATE		2.93 (-3.34, 7.11)	0.75 (-4.82, 4.42)	-2.63 (-8.14, 0.51)
	RR-CATE		109.3 (90.85, 123.11)	102.38 (86.68, 113.78)	91.67 (78.04, 101.58)
P(N3 $\to$ N3)	%	10.1 (4.16, 17.88)	8.86 (5.38, 12.2)	7.78 (4.87, 10.51)	6.32 (4.1, 8.32)
	CATE		-1.24 (-11.52, 3.46)	-2.32 (-11.89, 2.35)	-3.79 (-12.42, 0.84)
	RR-CATE		87.76 (36.17, 171.39)	77.02 (32.56, 146.38)	62.53 (28.65, 119.37)
P(REM $\to$ REM)	%	9.4 (5.48, 20.11)	7.26 (6.07, 10.79)	6.91 (5.9, 10.11)	6.33 (5.39, 9.11)
	CATE		-2.15 (-9.37, 1.95)	-2.5 (-10.57, 1.67)	-3.07 (-12.08, 1.21)
	RR-CATE		77.16 (47.39, 136.07)	73.44 (45.74, 131)	67.35 (41.12, 121.62)
P(W-fragmentation)	%	3.25 (1.81, 5.02)	3.82 (3.07, 4.97)	4.15 (3.42, 5.2)	4.65 (3.99, 5.67)
	CATE		0.57 (-0.43, 2.19)	0.9 (-0.08, 2.5)	**1.39*** (0.37, 2.97)
	RR-CATE		117.37 (90.6, 219.1)	127.59 (98.35, 238.16)	**142.83*** (107.28, 262.14)
P(N1-fragmentation)	%	6.12 (4.42, 8.02)	7.49 (6.58, 9.06)	8.19 (7.34, 9.74)	9.24 (8.57, 10.94)
	CATE		**1.37*** (0.05, 4.43)	**2.07*** (1.01, 5.04)	**3.12*** (2.08, 6)
	RR-CATE		**122.43*** (100.68, 199.22)	**133.8*** (112.74, 214.14)	**151.02*** (126.83, 234.56)
P(N2-fragmentation)	%	5.49 (3.49, 8.55)	7.5 (6.53, 9.65)	7.88 (7, 9.91)	8.45 (7.7, 10.22)
	CATE		**2.01*** (0.36, 4.42)	**2.4*** (0.75, 4.7)	**2.96*** (1.31, 5.11)
	RR-CATE		**136.67*** (104.35, 227.21)	**143.69*** (108.64, 234.71)	**154.01*** (115.91, 246.5)
P(N3-fragmentation)	%	2.21 (1.18, 3.3)	3.33 (2.77, 4.2)	3.35 (2.84, 4.09)	3.36 (2.86, 4)
	CATE		**1.13*** (0.39, 2.38)	**1.15*** (0.43, 2.28)	**1.15*** (0.46, 2.09)
	RR-CATE		**151.09*** (110.91, 297.74)	**152.13*** (115.04, 289.05)	**152.37*** (115.08, 273.77)
P(REM-fragmentation)	%	1.68 (0.8, 2.41)	2.41 (1.96, 3.12)	2.55 (2.07, 3.2)	2.76 (2.33, 3.3)
	CATE		**0.73*** (0.23, 1.79)	**0.87*** (0.36, 1.86)	**1.07*** (0.53, 1.97)
	RR-CATE		**143.11*** (110.63, 313.22)	**151.73*** (116.31, 328.02)	**163.67*** (123.05, 338.27)

Notes: Probabilities are expressed as percentages. Estimates include conditional average treatment effects (CATE) and risk-ratio CATE (RR-CATE).

B.4 Comparison based on derived Markovian matrices P^M

Figure B.7: Expected derived Markovian transition matrices $\mathbf{P}^M$ for healthy females and females with different OSA severities, each stratified by age.



Notes: OSA severities are represented by AHI levels of 5, 15, and 30, and ages by 30, 50, and 70 years. Estimates are shown with 95% bootstrapped confidence intervals (CIs).

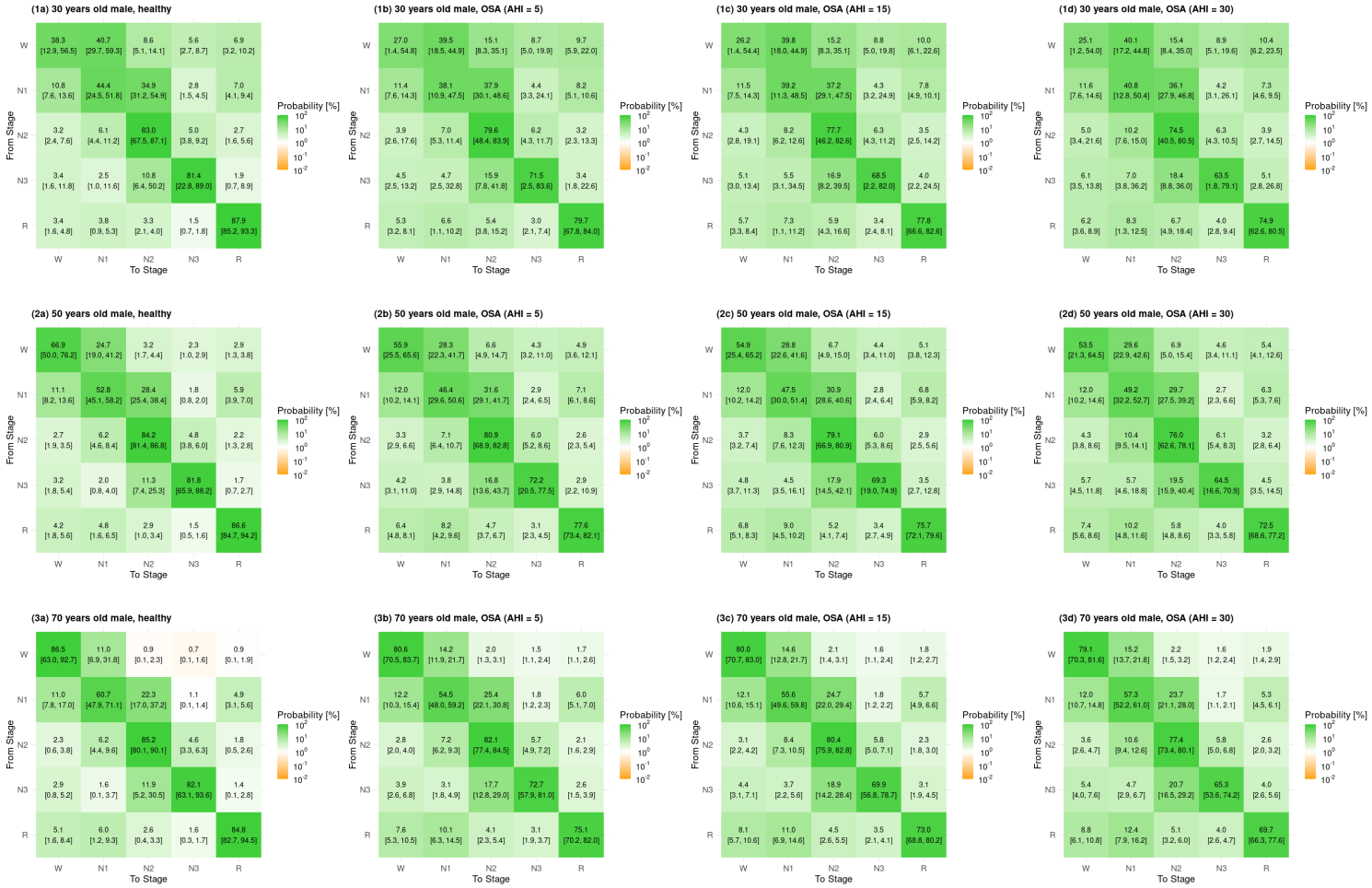
Figure B.8: Differences (CATE) in derived Markovian transition matrices $\mathbf{P}^M$ between healthy females and females with different OSA severities, each stratified by age.

Notes: OSA severities are represented by AHI levels of 5, 15, and 30, and ages by 30, 50, and 70 years. Estimates are shown with 95% bootstrapped confidence intervals (CIs).

Figure B.9: Risk ratio (RR-CATE) of derived Markovian transition matrices $\mathbf{P}^M$ between healthy females and females with different OSA severities, each stratified by age.

Notes: OSA severities are represented by AHI levels of 5, 15, and 30, and ages by 30, 50, and 70 years. Estimates are shown with 95% bootstrapped confidence intervals (CIs).

Figure B.10: Expected derived Markovian transition matrices $\mathbf{P}^M$ for healthy males and males with different OSA severities, each stratified by age.



Notes: OSA severities are represented by AHI levels of 5, 15, and 30, and ages by 30, 50, and 70 years. Estimates are shown with 95% bootstrapped confidence intervals (CIs).

Figure B.11: Differences (CATE) in derived Markovian transition matrices $\mathbf{P}^M$ between healthy males and males with different OSA severities, each stratified by age.

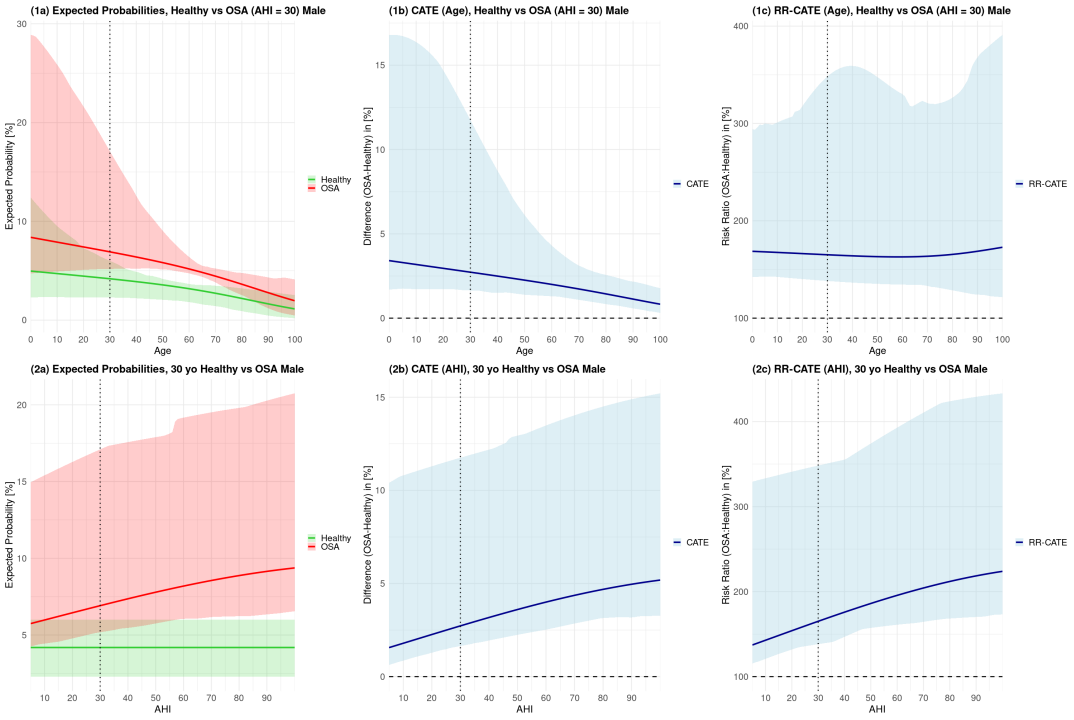
Notes: OSA severities are represented by AHI levels of 5, 15, and 30, and ages by 30, 50, and 70 years. Estimates are shown with 95% bootstrapped confidence intervals (CIs).

Figure B.12: Risk ratio (RR-CATE) of derived Markovian transition matrices $\mathbf{P}^M$ between healthy males and males with different OSA severities, each stratified by age.

Notes: OSA severities are represented by AHI levels of 5, 15, and 30, and ages by 30, 50, and 70 years. Estimates are shown with 95% bootstrapped confidence intervals (CIs).

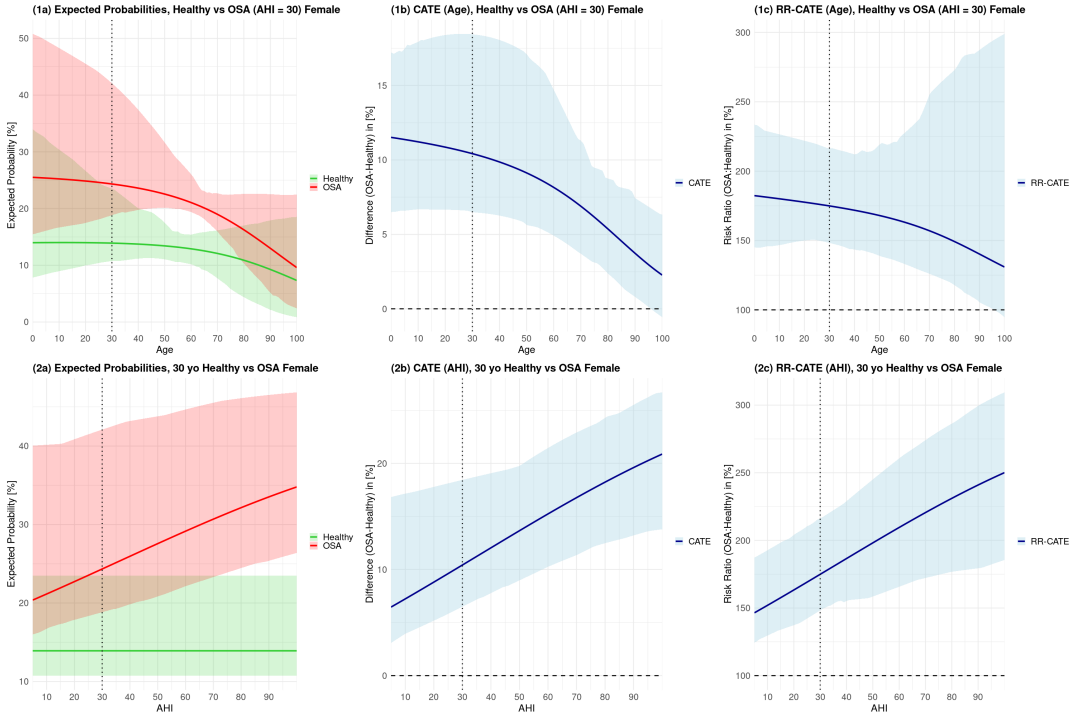
B.5 Effect plots for sample dynamics markers

Figure B.13: Effects of age and OSA-severities on NREM-REM oscillations, $P(\text{NREM} \rightleftarrows \text{REM})$, in males.



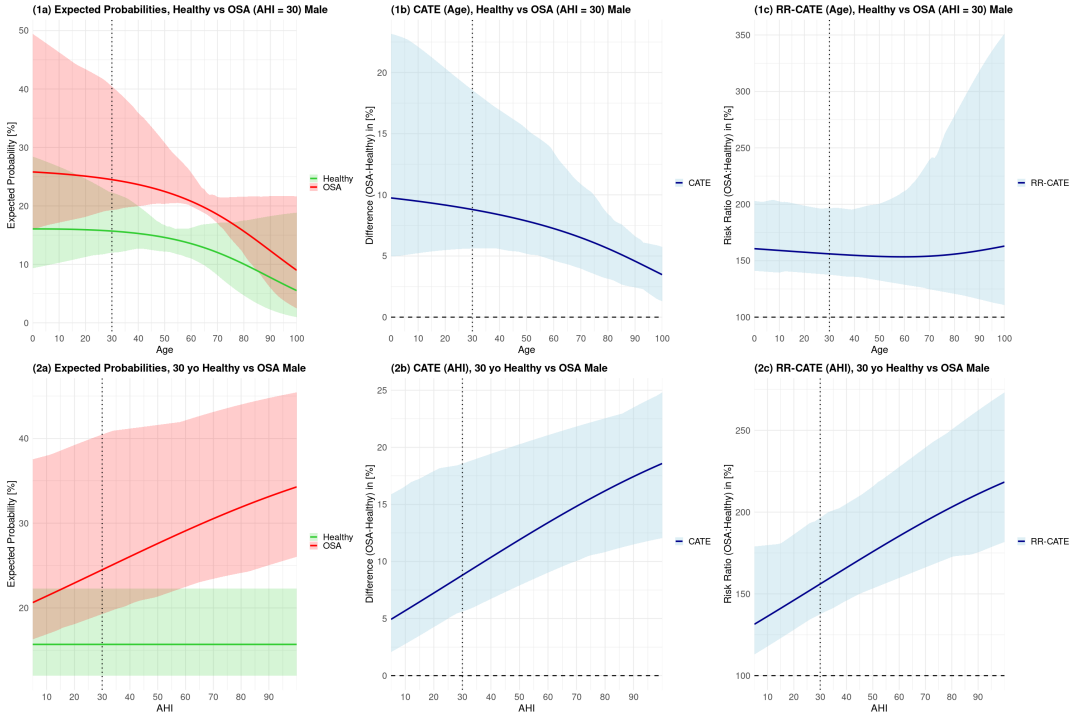
Notes: The left plots (1a, 2a) depict expected probabilities for varying age with fixed AHI = 30, and for varying AHI with fixed age = 30. Based on that, the central (1b, 2b) and right (1c, 2c) plots depict age- and AHI-related CATE and RR-CATE.

Figure B.14: Effects of age and OSA-severities on sleep-stage fragmentation, in females.



Notes: The sleep-stage fragmentation represents the probability of transitioning from one (non-wake) sleep stage to a different one, according to Eq. 5.12. The left plots (1a, 2a) depict expected probabilities for varying age with fixed AHI = 30, and for varying AHI with fixed age = 30. Based on that, the central (1b, 2b) and right (1c, 2c) plots depict age- and AHI-related CATE and RR-CATE.

Figure B.15: Effects of age and OSA-severities on sleep-stage fragmentation, in males.



Notes: The sleep-stage fragmentation represents the probability of transitioning from one (non-wake) sleep stage to a different one, according to Eq. 5.12. The left plots (1a, 2a) depict expected probabilities for varying age with fixed AHI = 30, and for varying AHI with fixed age = 30. Based on that, the central (1b, 2b) and right (1c, 2c) plots depict age- and AHI-related CATE and RR-CATE.

Appendix C

Supplementary Materials for Chapter 7

C.1 Bern Sleep-Wake Registery (BSWR)

C.1.1 Descriptive statistics

Table C.1: Characteristics of Berner Sleep-Wake Registry (BSWR) cohort stratified by no-to-mild SDB (AHI $\leq$ 15) versus moderate-to-high SDB (AHI > 15).

Variable	Overall	AHI $\leq$ 15	AHI > 15	p-value
N	3702	2100	1602	
Age	48.28 (19.37)	42.64 (19.53)	55.66 (16.46)	<0.001
Gender (Male)*	2325 (62.8)	1151 (54.8)	1174 (73.3)	<0.001
Smoking*				0.017
Current	179 (4.8)	99 (4.7)	80 (5.0)	
Ex	67 (1.8)	27 (1.3)	40 (2.5)	
Never	221 (6.0)	115 (5.5)	106 (6.6)	
NA	3235 (87.4)	1859 (88.5)	1376 (85.9)	
BMI	26.94 (6.46)	25.52 (6.23)	28.81 (6.29)	<0.001
AHI	19.02 (20.19)	6.31 (4.23)	35.69 (20.71)	<0.001
SDB (AHI>15)*	1602 (43.3)	0 (0.0)	1602 (100.0)	<0.001
SDB category*				<0.001
Mixed	951 (25.7)	0 (0.0)	951 (59.4)	
NREM-dominant	484 (13.1)	0 (0.0)	484 (30.2)	
REM-dominant	137 (3.7)	0 (0.0)	137 (8.6)	
AHI$\leq$15	2100 (56.7)	2100 (100.0)	0 (0.0)	
NA	30 (0.8)	0 (0.0)	30 (1.9)	
TST [mins]	339.13 (89.41)	354.88 (91.52)	318.49 (82.16)	<0.001
WASO [mins]	64.83 (54.42)	56.47 (49.68)	75.78 (58.30)	<0.001
SE [%]	80.03 (14.69)	82.42 (13.58)	76.89 (15.47)	<0.001
SL [mins]	18.64 (25.84)	18.20 (24.83)	19.22 (27.11)	0.235
REML [mins]	172.40 (186.46)	161.62 (169.84)	186.52 (205.42)	<0.001
DL [mins]	74.92 (185.01)	50.74 (137.39)	106.61 (229.36)	<0.001
W [%]	16.45 (13.74)	14.04 (12.54)	19.60 (14.58)	<0.001
N1 [%]	16.87 (10.92)	13.05 (7.77)	21.89 (12.34)	<0.001
N2 [%]	36.40 (12.46)	39.00 (11.46)	33.00 (12.89)	<0.001
N3 [%]	16.93 (10.68)	19.36 (10.70)	13.73 (9.78)	<0.001
REM [%]	13.35 (7.05)	14.55 (6.93)	11.79 (6.90)	<0.001

Notes: Continuous variables are reported as mean (SD) and compared using Welch's two-sample *t*-test. Categorical variables, denoted by superscript *, are reported as counts (percentages) and compared using the chi-squared test. When expected cell counts were less than 5, Fisher's exact test was used instead.

C.1.2 Occurrence of clinical conditions

Table C.2 shows the number and percentage, N (%), of PSG recordings in the BSWR, stratified by conclusive sleep diagnoses and non-sleep comorbidities and by the presence of sleep-disordered breathing (SDB): no-to-mild (AHI$\leq$15) versus moderate-to-severe (AHI>15). Out of 3702 recordings from 3417 unique subjects, 2100 had AHI$\leq$15 and 1602 had AHI>15. A total of 88 recordings corresponded to healthy individuals without any sleep diagnosis or on-sleep comorbidity, or undergoing PSG as healthy controls in some of the Inselspital's conducted clinical studies, without any clinical condition identified. The classification of major sleep disorder categories in the table is consistent with ICSD-3 (International Classification of Sleep Disorders, Third Edition), which defines seven major groups, such as SDB, Insomnias, and Central Disorders of Hypersomnolence (Hypersomnias). Within the most prevalent classes, we also considered the most relevant/common diagnoses subcategories (e.g., OSA and CSA within SDB, NT1 and NT2 within Hypersomnias) versus the less frequent "Other" conditions (e.g., hypoventilation and hypoxia syndrome in SDB). Rare categories (e.g., circadian rhythm disorders) and those with non-specific clinical profiles (e.g., isolated symptoms and normal variants) were grouped into a single main class.

We further grouped the available comorbidities into broader condition categories. The *Brain* category comprises individuals with a history of major neurological events, including stroke, intracerebral hemorrhage, and traumatic brain injury (TBI), as well as cases with suspected TBI-related diagnoses. This grouping captures subjects with structural brain injuries that may influence sleep physiology or contribute to comorbid conditions. The *Neurodegenerative* category comprises individuals with a confirmed or probable diagnosis of a neurodegenerative disorder. This includes Parkinson's disease, atypical Parkinsonian syndromes, amyotrophic lateral sclerosis (ALS), dementia, and other specified neurodegenerative conditions. The *Headache* category includes individuals with a confirmed or probable diagnosis of migraine, tension-type headache, post-traumatic headache, cluster headache, trigeminal neuralgia, and other specified headache syndromes. The *Psychiatric* category includes individuals with a confirmed or probable psychiatric disorder. This encompasses depression, bipolar disorder, anxiety and panic disorders, conversion disorder, post-traumatic stress disorder (PTSD), attention-deficit/hyperactivity disorder (ADHD), and substance-related disorders (alcohol and drug abuse), as well as other specified psychiatric conditions. The *Diabetes* category includes individuals with a confirmed or probable diagnosis of diabetes mellitus. The *Cardial* category includes probable or confirmed hypertension, coronary heart disease, atrial fibrillation, and other cardiac comorbidities. The *Pulmonary* category includes individuals with a confirmed or probable diagnosis of chronic respiratory disease, including chronic obstructive pulmonary disease (COPD), asthma, and other specified pulmonary conditions.

Table C.2: Health conditions in the Berner Sleep-Wake Registry (BSWR) stratified by no-to-mild SDB (AHI $\leq$ 15) versus moderate-to-high SDB (AHI > 15).

Diagnosis	N (%)	AHI $\leq$ 15	AHI>15	p-value
Total	3702	2100	1602	
Healthy	88 (2.4)	88 (4.2)	0 (0.0)	<0.001
Sleep Disorders:				
Sleep-disordered breathing (SDB)	2695 (72.8)	1120 (53.3)	1575 (98.3)	<0.001
Obstructive Sleep Apnea (OSA)	687 (18.6)	323 (15.4)	364 (22.7)	<0.001
Central Sleep Apnea (CSA)	111 (3.0)	33 (1.6)	78 (4.9)	<0.001
Other	1957 (52.9)	786 (37.4)	1171 (73.1)	<0.001
Insomnias	487 (13.2)	337 (16.0)	150 (9.4)	<0.001
Chronic	105 (2.8)	81 (3.9)	24 (1.5)	<0.001
Short-term	2 (0.1)	2 (0.1)	0 (0.0)	0.602
Other	385 (10.4)	257 (12.2)	128 (8.0)	<0.001
Hypersomnias	715 (19.3)	584 (27.8)	131 (8.2)	<0.001
Narcolepsy Type 1 (NT1)	58 (1.6)	36 (1.7)	22 (1.4)	0.488
Narcolepsy Type 2 (NT2)	8 (0.2)	7 (0.3)	1 (0.1)	0.161
Idiopathic Hypersomnia (IH)	24 (0.6)	24 (1.1)	0 (0.0)	<0.001
Excessive Daytime Sleepiness (EDS)	388 (10.5)	332 (15.8)	56 (3.5)	<0.001
Other	310 (8.4)	251 (12.0)	59 (3.7)	<0.001
Movement-related	265 (7.2)	166 (7.9)	99 (6.2)	0.051
Restless Leg Syndrome (RLS)	182 (4.9)	101 (4.8)	81 (5.1)	0.789
Periodic Limb Movement Disorder (PLMD)	34 (0.9)	27 (1.3)	7 (0.4)	0.012
Other	50 (1.4)	38 (1.8)	12 (0.7)	0.009
Parasomnias	193 (5.2)	124 (5.9)	69 (4.3)	0.036
REM	55 (1.5)	49 (2.3)	6 (0.4)	<0.001
NREM	128 (3.5)	68 (3.2)	60 (3.7)	0.456
Other	17 (0.5)	10 (0.5)	7 (0.4)	1.000
Circadian-rhythm-related	47 (1.3)	32 (1.5)	15 (0.9)	0.152
Isolated symptoms and norm variants	1771 (47.8)	994 (47.3)	777 (48.5)	0.502
Non-sleep Comorbidities:				
Brain	64 (1.7)	34 (1.6)	30 (1.9)	0.646
Neurodegenerative	81 (2.2)	47 (2.2)	34 (2.1)	0.900
Epilepsy	49 (1.3)	35 (1.7)	14 (0.9)	0.052
Headache	73 (2.0)	51 (2.4)	22 (1.4)	0.030
Psychiatric	204 (5.5)	145 (6.9)	59 (3.7)	<0.001
Diabetes	41 (1.1)	22 (1.0)	19 (1.2)	0.810
Cardial	134 (3.6)	54 (2.6)	80 (5.0)	<0.001
Pulmonary	50 (1.4)	35 (1.7)	15 (0.9)	0.078

Notes: Number and percentage (N, %) of different health conditions in BSWR, stratified by the presence of no-to-mild sleep-disordered breathing (SDB; AHI $\leq$ 15) versus moderate-to-severe SDB (AHI > 15). Equality of proportions between SDB groups was assessed using the chi-squared test or Fisher's exact test when expected cell counts were less than 5. Results are reported as p-values, with significant differences highlighted as follows: for $p < 0.05$, for $p < 0.01$, and for $p < 0.001$.

C.1.3 Characteristics of clinical conditions, predicted risk, and their comparison to healthy.

Table C.3: Summary statistics and adjusted cardiovascular risk in the Bern Sleep-Wake Registry (BSWR).

Diagnosis	Gender-Male	Age	BMI	AHI	Predicted Risk	Adjusted Risk [%]
Healthy	37 (42.0%)	32.86 (18.81)	22.63 (5.31)	2.68 (1.85)	17.14 (4.10)	NA
Sleep Disorders:						
Sleep-disordered breathing (SDB)	1868 (69.3%)***	51.1 (19.0)***	27.6 (6.5)***	24.6 (20.8)***	27.9 (9.9)***	17.8 (9.5, 26.8)***
Obstructive Sleep Apnea (OSA)	476 (69.3%)***	51.9 (17.1)***	27.8 (6.5)***	22.7 (19.8)***	27.1 (8.4)***	18.5 (9.4, 28.3)***
Central Sleep Apnea (CSA)	74 (66.7%)***	45.6 (24.7)***	25.0 (6.0)**	29.6 (22.9)***	28.0 (9.5)***	39.2 (23.5, 56.9)***
Other	1359 (69.4%)***	51.1 (19.3)***	27.7 (6.6)***	24.9 (20.8)***	28.2 (10.3)***	17.4 (8.8, 26.7)***
Insomnias	262 (53.8%)	53.6 (15.4)***	26.6 (5.3)***	13.8 (15.7)***	26.7 (9.5)***	12.4 (2.9, 22.7)**
Chronic	61 (58.1%)*	50.4 (15.8)***	26.2 (4.4)***	9.5 (9.4)***	24.4 (9.3)***	20.4 (5.6, 37.3)**
Short-term	2 (100.0%)	61.0 (5.7)*	25.5 (3.5)	5.4 (1.1)	27.2 (3.7)	41.4 (-10.5, 123.3)
Other	202 (52.5%)	54.5 (15.3)***	26.6 (5.5)***	15.0 (16.8)***	27.4 (9.8)***	14.7 (4.5, 25.9)**
Hypersomnias	346 (48.4%)	39.5 (15.5)**	26.6 (6.4)***	9.3 (13.1)***	21.1 (6.9)***	11.0 (3.7, 18.9)**
Narcolepsy Type 1 (NT1)	27 (46.6%)	37.3 (16.7)	27.6 (5.0)***	14.4 (16.1)***	23.1 (7.5)***	22.9 (8.8, 38.9)**
Narcolepsy Type 2 (NT2)	6 (75.0%)	29.8 (20.6)	23.9 (5.0)	7.4 (8.8)	23.3 (8.9)	12.7 (-11.8, 44.0)
Idiopathic Hypersomnia (IH)	2 (8.3%)**	26.3 (7.0)**	24.3 (3.8)	2.7 (2.0)	17.2 (3.2)	13.7 (-1.7, 31.4)
Excessive Daytime Sleepiness (EDS)	162 (52.3%)	39.3 (15.0)**	26.8 (6.8)***	9.5 (13.0)***	20.7 (6.3)***	9.7 (1.4, 18.8)*
Other	179 (46.1%)	40.1 (15.4)**	26.6 (6.7)***	8.2 (12.0)***	20.9 (7.1)***	10.8 (2.9, 19.2)**
Movement-related	164 (61.9%)**	53.2 (17.5)***	27.3 (6.4)***	16.5 (18.5)***	28.2 (11.1)***	20.3 (8.6, 33.2)***
Restless Leg Syndrome	103 (56.6%)*	58.2 (15.1)***	27.7 (6.5)***	19.3 (20.3)***	30.0 (11.8)***	23.4 (8.5, 40.4)**
Periodic Limb Movement Disorder	23 (67.6%)*	51.1 (17.7)***	26.2 (4.7)***	9.7 (9.8)***	27.2 (9.2)***	32.6 (13.2, 55.3)***
Other	38 (76.0%)***	37.2 (15.6)	26.8 (6.7)***	11.3 (13.3)***	22.6 (7.4)***	12.9 (-0.5, 28.2)
Parasomnias	128 (66.3%)***	53.9 (19.1)***	25.8 (5.1)***	14.3 (14.5)***	27.3 (8.5)***	24.5 (13.3, 36.9)***
REM	36 (65.5%)*	34.8 (16.1)	24.3 (4.0)*	6.9 (8.1)***	21.1 (5.3)***	16.1 (3.6, 30.2)*
NREM	83 (64.8%)**	62.9 (12.4)***	26.2 (5.2)***	17.8 (15.7)***	30.2 (7.8)***	44.4 (27.0, 64.1)***
Other	14 (82.4%)**	50.6 (20.8)**	26.3 (5.4)*	11.4 (8.3)***	28.2 (11.7)**	17.3 (-9.2, 51.5)
Circadian-rhythm-related	33 (70.2%)**	44.5 (17.1)***	27.6 (5.6)***	14.5 (15.2)***	26.8 (9.5)***	27.5 (10.3, 47.3)**
Isolated symptoms and norm variants	1162 (65.6%)***	52.0 (17.7)***	27.4 (6.0)***	18.6 (18.9)***	27.4 (10.4)***	10.9 (2.9, 19.5)**
Non-sleep Comorbidities:						
Brain	47 (73.4%)***	53.6 (14.2)***	27.4 (4.6)***	21.5 (21.8)***	25.6 (9.1)***	10.2 (-5.5, 28.5)
Neurodegenerative	45 (55.6%)	63.9 (9.7)***	26.2 (4.8)***	18.0 (17.8)***	29.9 (9.3)***	45.8 (21.4, 75.1)***
Epilepsy	35 (71.4%)**	44.7 (14.3)***	26.5 (5.7)***	11.4 (12.3)***	23.0 (6.3)***	14.5 (0.9, 30.0)*
Headache	36 (49.3%)	42.2 (15.5)***	27.7 (7.1)***	16.4 (21.8)***	22.7 (8.4)***	13.0 (0.6, 27.0)*
Psychiatric	105 (51.5%)	45.8 (16.2)***	28.1 (6.2)***	12.9 (15.4)***	23.6 (7.5)***	21.2 (10.1, 33.3)***
Diabetes	29 (70.7%)**	55.9 (14.3)***	30.0 (8.7)***	25.7 (25.5)***	28.6 (9.3)***	39.4 (13.9, 70.7)**
Cardial	108 (80.6%)***	57.2 (12.4)***	30.7 (6.3)***	27.8 (24.2)***	29.5 (9.0)***	25.1 (8.7, 44.0)**
Pulmonary	33 (66.0%)*	46.5 (16.7)***	29.3 (8.0)***	15.7 (20.1)***	23.8 (9.5)***	19.0 (2.3, 38.3)*

Notes: The table reports Gender (N (%) of males) and mean (standard deviation) for Age, Body Mass Index (BMI), Apnea-Hypopnea Index (AHI), and predicted cardiovascular risk (mortality) in Bern Sleep-Wake Registry (BSWR). The adjusted risk, reported as an estimate (95% CI), quantifies the systematic percentual difference in predicted risk for specific diagnoses (conclusive sleep disorders and non-sleep comorbidities) using logistic regression adjusting for gender, age, BMI, and AHI. Significant differences in comparison to healthy controls are highlighted as: * if p-val$<$0.05, ** if p-val$<$0.01, and *** if p-val$<$0.001.

C.2 Sleep Heart Health Study (SHHS)

Table C.4: Subsets of SHHS database stratified based on prior cardiovascular event status (E) and medication use (M).

Dataset	N	N-events	Age	Gender
SHHS1 (E = 0, M = 0)	2579	326	59.43 (11.16)	1182 (45.8)
SHHS1 (E = 0, M = 1)	2528	567	64.97 (10.15)	1157 (45.8)
SHHS1 (E = 1, M = 0)	112	60	68.61 (11.86)	67 (59.8)
SHHS1 (E = 1, M = 1)	572	320	70.67 (9.68)	354 (61.9)
SHHS2 (E = 0, M = 0)	811	62	63.16 (10.49)	358 (44.1)
SHHS2 (E = 0, M = 1)	1484	201	68.73 (9.60)	647 (43.6)
SHHS2 (E = 1, M = 0)	37	15	70.97 (10.55)	22 (59.5)
SHHS2 (E = 1, M = 1)	319	178	73.52 (9.06)	199 (62.4)

Notes: Summary statistics include the number of subjects (N), the number who developed a cardiovascular event during follow-up (N-events), the mean (SD) of age in years, and the number (%) of males.

C.2.1 SHHS1

Table C.5: Descriptive characteristics of SHHS1 (E = 0, M = 1) cohort stratified by cardiovascular event status.

Variable	Overall	Event-free	Event developed	p-value
N	2528	1961	567	
Age	64.97 (10.15)	63.32 (9.97)	70.68 (8.57)	<0.001
Gender (Male)*	1157 (45.8)	859 (43.8)	298 (52.6)	<0.001
Smoking*				<0.001
Current	221 (8.7)	166 (8.5)	55 (9.7)	
Ex	1121 (44.3)	831 (42.4)	290 (51.1)	
Never	1174 (46.4)	954 (48.6)	220 (38.8)	
NA	12 (0.5)	10 (0.5)	2 (0.4)	
BMI	28.64 (5.22)	28.60 (5.24)	28.77 (5.14)	0.507
AHI	18.89 (16.90)	18.27 (16.85)	21.05 (16.90)	0.001
SDB (AHI>15)*	1179 (46.6)	863 (44.0)	316 (55.7)	<0.001
SDB category*				<0.001
Mixed	574 (22.7)	425 (21.7)	149 (26.3)	
NREM-dominant	108 (4.3)	78 (4.0)	30 (5.3)	
REM-dominant	379 (15.0)	271 (13.8)	108 (19.0)	
AHI≤15	1349 (53.4)	1098 (56.0)	251 (44.3)	
NA	118 (4.7)	89 (4.5)	29 (5.1)	
TST [mins]	358.28 (63.24)	359.36 (63.05)	354.54 (63.82)	0.110
WASO [mins]	95.97 (55.62)	93.97 (54.94)	102.90 (57.42)	0.001
SE [%]	70.89 (12.11)	71.06 (12.11)	70.28 (12.11)	0.175
SL [mins]	52.19 (42.78)	53.47 (43.32)	47.75 (40.56)	0.005
REML [mins]	122.90 (169.81)	122.04 (167.02)	125.89 (179.24)	0.634
DL [mins]	83.39 (228.77)	81.04 (225.39)	91.53 (240.14)	0.336
W [%]	20.98 (11.79)	20.58 (11.67)	22.38 (12.09)	0.001
N1 [%]	4.16 (2.85)	4.13 (2.81)	4.25 (3.00)	0.383
N2 [%]	45.87 (12.38)	45.80 (12.44)	46.09 (12.19)	0.622
N3 [%]	13.95 (9.99)	14.22 (9.98)	13.00 (9.98)	0.010
REM [%]	15.04 (6.27)	15.26 (6.23)	14.28 (6.35)	0.001

Notes: Continuous variables are reported as mean (SD) and compared using Welch's two-sample *t*-test. Categorical variables, denoted by superscript *, are reported as counts (percentages) and compared using the chi-squared test. When expected cell counts were less than 5, Fisher's exact test was used instead.

Table C.6: Descriptive characteristics of SHHS1 (E = 1, M = 0) cohort stratified by cardiovascular event status.

Variable	Overall	Event-free	Event developed	p-value
N	112	52	60	
Age	68.61 (11.86)	64.25 (13.48)	72.38 (8.74)	<0.001
Gender (Male)*	67 (59.8)	26 (50.0)	41 (68.3)	0.075
Smoking*				0.401
Current	10 (8.9)	5 (9.6)	5 (8.3)	
Ex	55 (49.1)	22 (42.3)	33 (55.0)	
Never	47 (42.0)	25 (48.1)	22 (36.7)	
BMI	27.80 (4.96)	27.73 (6.13)	27.86 (3.72)	0.896
AHI	20.99 (15.27)	18.97 (12.37)	22.74 (17.31)	0.194
SDB (AHI>15)*	67 (59.8)	30 (57.7)	37 (61.7)	0.814
SDB category*				0.524
Mixed	26 (23.2)	9 (17.3)	17 (28.3)	
NREM-dominant	9 (8.0)	4 (7.7)	5 (8.3)	
REM-dominant	24 (21.4)	14 (26.9)	10 (16.7)	
AHI≤15	45 (40.2)	22 (42.3)	23 (38.3)	
NA	8 (7.1)	3 (5.8)	5 (8.3)	
TST [mins]	349.95 (73.14)	370.17 (57.88)	332.42 (80.58)	0.006
WASO [mins]	101.90 (64.26)	81.72 (46.60)	119.39 (72.25)	0.002
SE [%]	69.66 (13.70)	73.91 (10.40)	65.98 (15.16)	0.002
SL [mins]	50.80 (41.73)	49.58 (35.41)	51.86 (46.79)	0.774
REML [mins]	124.81 (196.57)	126.16 (185.06)	123.63 (207.58)	0.946
DL [mins]	81.23 (222.33)	44.66 (139.63)	112.92 (271.93)	0.105
W [%]	22.47 (13.87)	18.03 (10.05)	26.32 (15.55)	0.001
N1 [%]	4.72 (3.27)	4.22 (2.96)	5.15 (3.49)	0.134
N2 [%]	45.44 (13.11)	47.31 (12.40)	43.82 (13.58)	0.160
N3 [%]	12.33 (9.73)	14.30 (11.16)	10.63 (8.01)	0.046
REM [%]	15.04 (6.39)	16.14 (5.68)	14.08 (6.84)	0.089

Notes: Continuous variables are reported as mean (SD) and compared using Welch's two-sample *t*-test. Categorical variables, denoted by superscript *, are reported as counts (percentages) and compared using the chi-squared test. When expected cell counts were less than 5, Fisher's exact test was used instead.

Table C.7: Descriptive characteristics of SHHS1 (E = 1, M = 1) cohort stratified by cardiovascular event status.

Variable	Overall	Event-free	Event developed	p-value
N	572	252	320	
Age	70.67 (9.68)	68.50 (10.28)	72.38 (8.83)	<0.001
Gender (Male)*	354 (61.9)	153 (60.7)	201 (62.8)	0.670
Smoking*				0.557
Current	46 (8.0)	21 (8.3)	25 (7.8)	
Ex	305 (53.3)	128 (50.8)	177 (55.3)	
Never	221 (38.6)	103 (40.9)	118 (36.9)	
BMI	28.08 (5.14)	27.81 (5.01)	28.29 (5.23)	0.271
AHI	22.06 (17.36)	20.23 (16.35)	23.50 (18.02)	0.025
SDB (AHI>15)*	323 (56.5)	131 (52.0)	192 (60.0)	0.067
SDB category*				0.296
Mixed	164 (28.7)	62 (24.6)	102 (31.9)	
NREM-dominant	34 (5.9)	14 (5.6)	20 (6.2)	
REM-dominant	82 (14.3)	37 (14.7)	45 (14.1)	
AHI$\leq$15	249 (43.5)	121 (48.0)	128 (40.0)	
NA	43 (7.5)	18 (7.1)	25 (7.8)	
TST [mins]	347.37 (69.96)	354.70 (66.21)	341.59 (72.37)	0.026
WASO [mins]	101.47 (58.91)	96.46 (56.00)	105.42 (60.91)	0.071
SE [%]	68.79 (13.51)	70.26 (13.25)	67.63 (13.61)	0.021
SL [mins]	57.04 (49.31)	55.19 (47.87)	58.49 (50.44)	0.427
REML [mins]	130.81 (202.14)	114.44 (170.03)	143.70 (223.64)	0.086
DL [mins]	117.64 (274.90)	90.64 (232.66)	138.89 (302.73)	0.037
W [%]	22.58 (12.94)	21.30 (12.07)	23.59 (13.51)	0.035
N1 [%]	4.38 (3.22)	4.38 (2.90)	4.38 (3.46)	0.995
N2 [%]	46.09 (13.68)	46.30 (12.71)	45.92 (14.42)	0.745
N3 [%]	12.58 (10.50)	12.94 (9.68)	12.30 (11.10)	0.463
REM [%]	14.37 (6.43)	15.08 (6.31)	13.81 (6.47)	0.019

Notes: Continuous variables are reported as mean (SD) and compared using Welch's two-sample *t*-test. Categorical variables, denoted by superscript *, are reported as counts (percentages) and compared using the chi-squared test. When expected cell counts were less than 5, Fisher's exact test was used instead.

C.2.2 SHHS2

Table C.8: Descriptive characteristics of SHHS2 (E = 0, M = 0) cohort stratified by cardiovascular event status.

Variable	Overall	Event-free	Event developed	p-value
N	811	749	62	
Age	63.16 (10.49)	62.52 (10.16)	70.89 (11.39)	<0.001
Gender (Male)*	358 (44.1)	321 (42.9)	37 (59.7)	0.015
Smoking*				0.010
Current	62 (7.6)	51 (6.8)	11 (17.7)	
Ex	296 (36.5)	272 (36.3)	24 (38.7)	
Never	445 (54.9)	418 (55.8)	27 (43.5)	
NA	8 (1.0)	8 (1.1)	0 (0.0)	
BMI	27.72 (4.69)	27.70 (4.69)	27.90 (4.73)	0.749
AHI	15.84 (15.38)	15.46 (15.16)	20.35 (17.33)	0.016
SDB (AHI>15)*	300 (37.0)	268 (35.8)	32 (51.6)	0.019
SDB category*				0.003
Mixed	158 (19.5)	143 (19.1)	15 (24.2)	
NREM-dominant	20 (2.5)	18 (2.4)	2 (3.2)	
REM-dominant	106 (13.1)	96 (12.8)	10 (16.1)	
AHI≤15	511 (63.0)	481 (64.2)	30 (48.4)	
NA	16 (2.0)	11 (1.5)	5 (8.1)	
TST [mins]	384.11 (60.69)	385.78 (59.02)	363.86 (75.76)	0.006
WASO [mins]	155.78 (65.97)	154.41 (65.70)	172.35 (67.59)	0.040
SE [%]	64.53 (10.36)	64.74 (10.28)	62.03 (11.10)	0.047
SL [mins]	59.44 (39.19)	60.15 (39.28)	50.88 (37.36)	0.073
REML [mins]	99.65 (107.53)	97.15 (100.56)	129.87 (168.91)	0.021
DL [mins]	48.33 (132.58)	47.30 (132.59)	60.79 (132.83)	0.442
W [%]	28.45 (10.67)	28.16 (10.53)	32.03 (11.75)	0.006
N1 [%]	3.79 (3.76)	3.72 (3.80)	4.58 (3.09)	0.084
N2 [%]	40.38 (9.28)	40.35 (9.30)	40.79 (9.05)	0.715
N3 [%]	11.95 (7.86)	12.18 (7.88)	9.19 (7.19)	0.004
REM [%]	15.43 (5.40)	15.60 (5.39)	13.41 (5.18)	0.002

Notes: Continuous variables are reported as mean (SD) and compared using Welch's two-sample *t*-test. Categorical variables, denoted by superscript *, are reported as counts (percentages) and compared using the chi-squared test. When expected cell counts were less than 5, Fisher's exact test was used instead.

Table C.9: Descriptive characteristics of SHHS2 (E = 0, M = 1) cohort stratified by cardiovascular event status.

Variable	Overall	Event-free	Event developed	p-value
N	1484	1283	201	
Age	68.73 (9.60)	67.95 (9.51)	73.72 (8.68)	<0.001
Gender (Male)*	647 (43.6)	537 (41.9)	110 (54.7)	0.001
Smoking*				0.082
Current	101 (6.8)	80 (6.2)	21 (10.4)	
Ex	644 (43.4)	552 (43.0)	92 (45.8)	
Never	719 (48.5)	634 (49.4)	85 (42.3)	
NA	20 (1.3)	17 (1.3)	3 (1.5)	
BMI	28.72 (5.30)	28.79 (5.32)	28.26 (5.19)	0.190
AHI	18.60 (16.29)	18.18 (16.17)	21.31 (16.85)	0.011
SDB (AHI>15)*	680 (45.8)	569 (44.3)	111 (55.2)	0.005
SDB category*				0.036
Mixed	376 (25.3)	310 (24.2)	66 (32.8)	
NREM-dominant	52 (3.5)	43 (3.4)	9 (4.5)	
REM-dominant	226 (15.2)	195 (15.2)	31 (15.4)	
AHI$\leq$15	804 (54.2)	714 (55.7)	90 (44.8)	
NA	26 (1.8)	21 (1.6)	5 (2.5)	
TST [mins]	373.67 (71.17)	376.93 (70.01)	352.85 (75.07)	<0.001
WASO [mins]	166.48 (74.83)	163.83 (73.64)	183.40 (80.18)	0.001
SE [%]	62.26 (11.95)	62.72 (11.78)	59.34 (12.59)	<0.001
SL [mins]	64.18 (46.50)	64.55 (46.24)	61.85 (48.15)	0.445
REML [mins]	118.48 (131.83)	114.62 (119.67)	143.11 (190.70)	0.004
DL [mins]	67.58 (182.98)	62.35 (173.68)	100.95 (231.55)	0.005
W [%]	30.44 (12.33)	29.91 (12.04)	33.81 (13.60)	<0.001
N1 [%]	3.88 (3.05)	3.79 (2.41)	4.43 (5.61)	0.006
N2 [%]	40.32 (10.44)	40.51 (10.36)	39.10 (10.89)	0.077
N3 [%]	11.03 (8.09)	11.20 (8.09)	9.90 (8.02)	0.033
REM [%]	14.34 (5.69)	14.58 (5.60)	12.75 (6.04)	<0.001

Notes: Continuous variables are reported as mean (SD) and compared using Welch's two-sample *t*-test. Categorical variables, denoted by superscript *, are reported as counts (percentages) and compared using the chi-squared test. When expected cell counts were less than 5, Fisher's exact test was used instead.

Table C.10: Descriptive characteristics of SHHS2 (E = 1, M = 0) cohort stratified by cardiovascular event status.

Variable	Overall	Event-free	Event developed	p-value
N	37	22	15	
Age	70.97 (10.55)	69.45 (12.55)	73.20 (6.43)	0.296
Gender (Male)*	22 (59.5)	11 (50.0)	11 (73.3)	0.281
Smoking*				0.242
Current	2 (5.4)	2 (9.1)	0 (0.0)	
Ex	17 (45.9)	8 (36.4)	9 (60.0)	
Never	18 (48.6)	12 (54.5)	6 (40.0)	
BMI	27.74 (5.06)	27.34 (5.73)	28.34 (3.98)	0.562
AHI	24.87 (16.70)	24.80 (19.32)	24.96 (12.56)	0.979
SDB (AHI>15)*	27 (73.0)	15 (68.2)	12 (80.0)	0.676
SDB category*				0.689
Mixed	15 (40.5)	7 (31.8)	8 (53.3)	
NREM-dominant	3 (8.1)	2 (9.1)	1 (6.7)	
REM-dominant	8 (21.6)	5 (22.7)	3 (20.0)	
AHI$\leq$15	10 (27.0)	7 (31.8)	3 (20.0)	
NA	1 (2.7)	1 (4.5)	0 (0.0)	
TST [mins]	362.89 (65.95)	358.98 (75.93)	368.63 (49.76)	0.668
WASO [mins]	182.27 (74.72)	188.80 (89.27)	172.70 (47.33)	0.528
SE [%]	59.95 (10.58)	59.23 (12.05)	61.00 (8.24)	0.623
SL [mins]	65.80 (41.20)	65.36 (42.79)	66.43 (40.20)	0.939
REML [mins]	78.95 (50.29)	75.27 (50.41)	84.33 (51.38)	0.598
DL [mins]	45.00 (62.72)	37.82 (37.15)	55.53 (88.58)	0.407
W [%]	32.88 (11.30)	33.72 (13.45)	31.66 (7.39)	0.594
N1 [%]	3.76 (2.34)	3.87 (2.67)	3.60 (1.84)	0.739
N2 [%]	39.74 (9.57)	39.09 (10.67)	40.70 (7.96)	0.623
N3 [%]	10.38 (7.23)	9.20 (6.12)	12.12 (8.52)	0.232
REM [%]	13.23 (5.87)	14.13 (6.92)	11.92 (3.69)	0.266

Notes: Continuous variables are reported as mean (SD) and compared using Welch's two-sample *t*-test. Categorical variables, denoted by superscript *, are reported as counts (percentages) and compared using the chi-squared test. When expected cell counts were less than 5, Fisher's exact test was used instead.

Table C.11: Descriptive characteristics of SHHS2 (E = 1, M = 1) cohort stratified by cardiovascular event status.

Variable	Overall	Event-free	Event developed	p-value
N	319	141	178	
Age	73.52 (9.06)	71.02 (9.87)	75.50 (7.84)	<0.001
Gender (Male)*	199 (62.4)	92 (65.2)	107 (60.1)	0.410
Smoking*				0.471
Current	24 (7.5)	13 (9.2)	11 (6.2)	
Ex	161 (50.5)	73 (51.8)	88 (49.4)	
Never	129 (40.4)	54 (38.3)	75 (42.1)	
NA	5 (1.6)	1 (0.7)	4 (2.2)	
BMI	28.03 (4.54)	28.00 (4.38)	28.05 (4.67)	0.922
AHI	23.56 (17.73)	23.25 (17.40)	23.80 (18.03)	0.784
SDB (AHI>15)*	194 (60.8)	90 (63.8)	104 (58.4)	0.386
SDB category*				0.822
Mixed	122 (38.2)	55 (39.0)	67 (37.6)	
NREM-dominant	14 (4.4)	6 (4.3)	8 (4.5)	
REM-dominant	45 (14.1)	23 (16.3)	22 (12.4)	
AHI$\leq$15	125 (39.2)	51 (36.2)	74 (41.6)	
NA	13 (4.1)	6 (4.3)	7 (3.9)	
TST [mins]	353.08 (76.90)	354.40 (77.68)	352.03 (76.48)	0.785
WASO [mins]	182.43 (79.74)	187.70 (86.63)	178.26 (73.81)	0.295
SE [%]	59.34 (12.32)	59.10 (13.26)	59.54 (11.56)	0.752
SL [mins]	62.49 (42.57)	62.05 (42.98)	62.84 (42.35)	0.869
REML [mins]	124.07 (164.14)	113.00 (144.40)	132.83 (178.15)	0.285
DL [mins]	105.08 (248.30)	102.35 (251.51)	107.23 (246.42)	0.862
W [%]	33.69 (13.32)	34.11 (14.35)	33.36 (12.47)	0.618
N1 [%]	4.41 (3.13)	4.17 (2.74)	4.61 (3.40)	0.215
N2 [%]	40.18 (11.10)	39.96 (10.95)	40.36 (11.25)	0.750
N3 [%]	9.09 (7.72)	8.90 (7.91)	9.24 (7.58)	0.696
REM [%]	12.62 (5.59)	12.86 (5.70)	12.43 (5.51)	0.497

Notes: Continuous variables are reported as mean (SD) and compared using Welch's two-sample *t*-test. Categorical variables, denoted by superscript *, are reported as counts (percentages) and compared using the chi-squared test. When expected cell counts were less than 5, Fisher's exact test was used instead.

C.3 Performance of Random Survival Forest model

Table C.12: Performance of the Random Survival Forest (RSF) model without AHI predictor across SHHS and BSWR datasets of subjects with no previous cardiovascular events (E = 0).

Metric	SHHS1^{CV} (E = 0, M = 0)	SHHS1† (E = 0, M = 1)	SHHS2 (E = 0, M = 0)	SHHS2† (E = 0, M = 0)	SHHS2 (E = 0, M = 1)	SHHS2† (E = 0, M = 1)
Events (N)	65.2 (2.4)	567	43	19	64	137
Event-free (N)	450.6 (2.7)	1961	591	158	489	794
C-index	73.3 (2.5)	69.7	74.1	78.3	70.1	66.6
IBS	6.7 (0.4)	11.9	4.1	6.2	7.1	8.8
1-year tdAUROC	77.1 (12.2)	69.8	85.2	77.1	72	64.1
5-year tdAUROC	74.9 (6.2)	72.7	73.6	83.4	71.4	69.2
10-year tdAUROC	75.3 (2.3)	74.2	-	100	-	69
Mortality (E = 1)	21.1 (1.8)	23.5	24.9	34.5	29.2	31.7
Mortality (E = 0)	11.3 (0.5)	14.5	14.1	16	17.9	21.6
Mortality Diff.	9.8 (1.8)	8.9	10.7	18.5	11.3	10.1
Mortality Diff. CI-low	5.8 (1.3)	7.6	4.7	9.1	6.4	6.6
Mortality Diff. CI-high	13.7 (2.4)	10.3	16.8	28	16.2	13.6
p-value (t-test)	0.000013 (0.000014)	$< 10^{-6}$	0.000856	0.000587	0.000019	$< 10^{-6}$
Events (N), high-risk	50.8 (1.9)	405	36	16	48	95
Events (N), low-risk	14.4 (1.9)	162	7	3	16	42
χ^2	26 (4.7)	162.5	22.1	10.4	20	28.4
p-value (log-rank test)	0.000005 (0.00001)	$< 10^{-6}$	0.000003	0.001237	0.000008	$< 10^{-6}$

Notes: The ^{CV} superscript denotes performance obtained via 5-fold cross-validation (CV) on the in-domain event- and medication-free (E = M = 0) baseline cohort SHHS1(E = 0, M = 0). All other columns evaluate the performance of the final RSF model fitted to the entire baseline cohort, applied to potentially out-of-domain subjects (†) from either the baseline (SHHS1) or follow-up (SHHS2) studies, including subgroups taking medication (M = 1). For each scenario, the number of subjects with events and without events (event-free) is reported. Model performance is assessed using Harrell's Concordance Index (C-index), Integrated Brier Score (IBS), and the time-dependent Area Under the Receiver Operating Characteristic curve (tdAUROC) at 1, 5, and 10 years. Discriminatory ability is evaluated via two-sided t-tests comparing predicted mortality between event and non-event subjects, including 95% confidence intervals (CI) for the difference (Diff.). Additionally, log-rank tests with Chi-squared (χ^2) statistics compare event rates between high- and low-risk groups stratified by median predicted mortality. For the in-domain CV, mean (SD) of all metrics is reported.

Table C.13: Performance of the Random Survival Forest model including AHI predictor across SHHS and BSWR datasets of subjects with previous cardiovascular events (E = 1).

Metric	SHHS1† (E = 1, M = 0)	SHHS1† (E = 1, M = 1)	SHHS2 (E = 1, M = 0)	SHHS2† (E = 1, M = 0)	SHHS2 (E = 1, M = 1)	SHHS2† (E = 1, M = 1)
Events (N)	60	320	2	13	15	163
Event-free (N)	52	252	0	22	0	141
C-index	60.8	62.6	0.0	56.6	62.9	60.8
IBS	27.4	28.5	22	22.6	32.8	28.9
1-year tdAUROC	29.2	64.9	-	78.1	100.0	62.5
5-year tdAUROC	66.6	67.4	-	57.2	91.7	65.5
10-year tdAUROC	73.2	69.3	-	-	-	-
Mortality (E = 1)	25.1	27.3	25.2	25	29.8	34.6
Mortality (E = 0)	18.1	21.0	-	25.4	-	27.1
Mortality Diff.	7.0	6.3	-	-0.4	-	7.5
Mortality Diff. CI-low	1.2	3.6	-	-11.5	-	3.1
Mortality Diff. CI-high	12.7	9.0	-	10.7	-	11.9
p-value (t-test)	0.017856	0.000005	-	0.94043	-	0.000846
Events (N), high-risk	36	184	1	7	7	95
Events (N), low-risk	24	136	1	6	8	68
χ^2	9.2	33.1	1.0	0.6	2.6	16.3
p-value (log-rank test)	0.002405	$< 10^{-6}$	0.317311	0.442614	0.106251	0.000053

Notes: The columns evaluate the performance of the final RSF model fitted to the entire event- and medication-free (E = M = 0) baseline cohort SHHS1(E = 0, M = 0), applied to potentially out-of-domain subjects (†) from either the baseline (SHHS1) or follow-up (SHHS2) studies, including subgroups taking medication (M = 1). For each scenario, the number of subjects with events and without events (event-free) is reported. Model performance is assessed using Harrell's Concordance Index (C-index), Integrated Brier Score (IBS), and the time-dependent Area Under the Receiver Operating Characteristic curve (tdAUROC) at 1, 5, and 10 years. Discriminatory ability is evaluated via two-sided t-tests comparing predicted mortality between event and non-event subjects, including 95% confidence intervals (CI) for the difference (Diff.). Additionally, log-rank tests with Chi-squared (χ^2) statistics compare event rates between high- and low-risk groups stratified by median predicted mortality.

Table C.14: Performance of the Random Survival Forest model without AHI predictor across SHHS and BSWR datasets of subjects with previous cardiovascular events (E = 1).

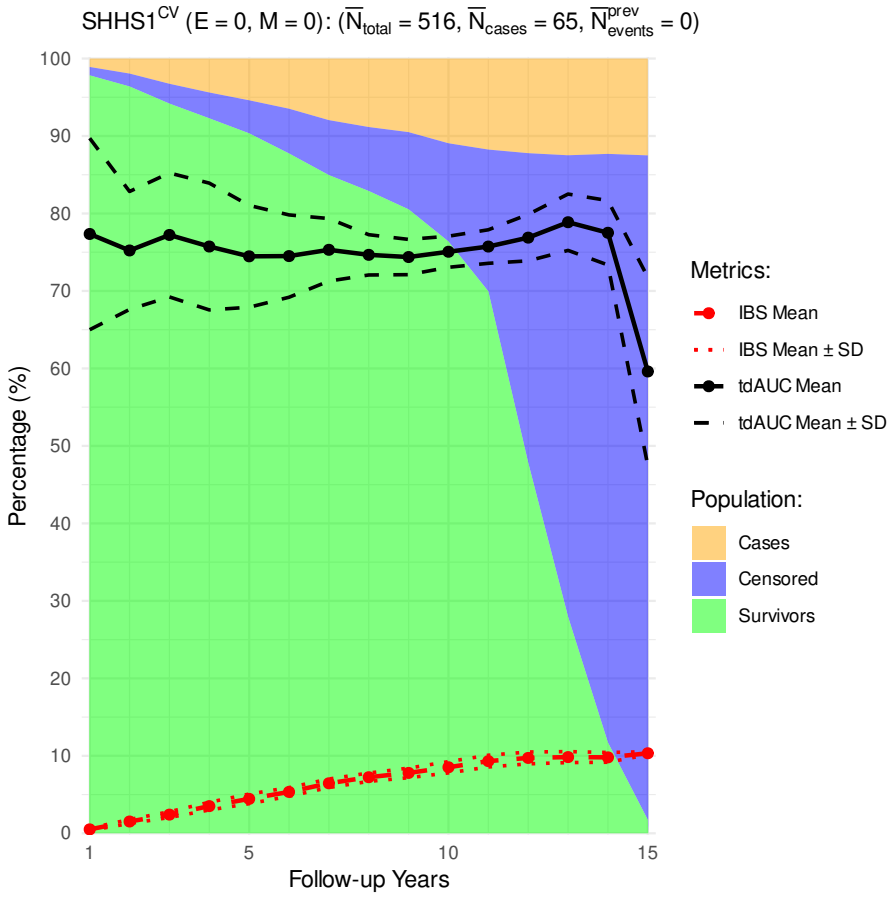
Metric	SHHS1† (E = 1, M = 0)	SHHS1† (E = 1, M = 1)	SHHS2 (E = 1, M = 0)	SHHS2† (E = 1, M = 0)	SHHS2 (E = 1, M = 1)	SHHS2† (E = 1, M = 1)
Events (N)	60	320	2	13	15	163
Event-free (N)	52	252	0	22	0	141
C-index	61.7	62.2	0	56.6	61.9	61.1
IBS	27.4	28.5	21.8	22.4	32.8	28.8
1-year tdAUROC	30.1	64.4	-	79.2	100	63
5-year tdAUROC	67.8	67	-	57.4	91.7	66
10-year tdAUROC	73.7	69.1	-	-	-	-
Mortality (E = 1)	25.0	27.3	25.1	25.5	30.1	34.6
Mortality (E = 0)	18.0	21.0	-	25.5	-	26.9
Mortality Diff.	7.0	6.3	-	0	-	7.7
Mortality Diff. CI-low	1.1	3.6	-	-11.3	-	3.3
Mortality Diff. CI-high	12.8	9.0	-	11.3	-	12.1
p-value (t-test)	0.019666	0.000006	-	0.997352	-	0.000583
Events (N), high-risk	36	181	1	7	7	95
Events (N), low-risk	24	139	1	6	8	68
χ^2	9.2	30.4	1.0	0.6	1.1	16.1
p-value (log-rank test)	0.002405	$< 10^{-6}$	0.317311	0.442614	0.286982	0.000061

Notes: The columns evaluate the performance of the final RSF model fitted to the entire event- and medication-free (E = M = 0) baseline cohort SHHS1(E = 0, M = 0), applied to potentially out-of-domain subjects (†) from either the baseline (SHHS1) or follow-up (SHHS2) studies, including subgroups taking medication (M = 1). For each scenario, the number of subjects with events and without events (event-free) is reported. Model performance is assessed using Harrell's Concordance Index (C-index), Integrated Brier Score (IBS), and the time-dependent Area Under the Receiver Operating Characteristic curve (tdAUROC) at 1, 5, and 10 years. Discriminatory ability is evaluated via two-sided t-tests comparing predicted mortality between event and non-event subjects, including 95% confidence intervals (CI) for the difference (Diff.). Additionally, log-rank tests with Chi-squared (χ^2) statistics compare event rates between high- and low-risk groups stratified by median predicted mortality.

C.4 Survival Plots for RSF with AHI predictor

C.4.1 Primary study cohort

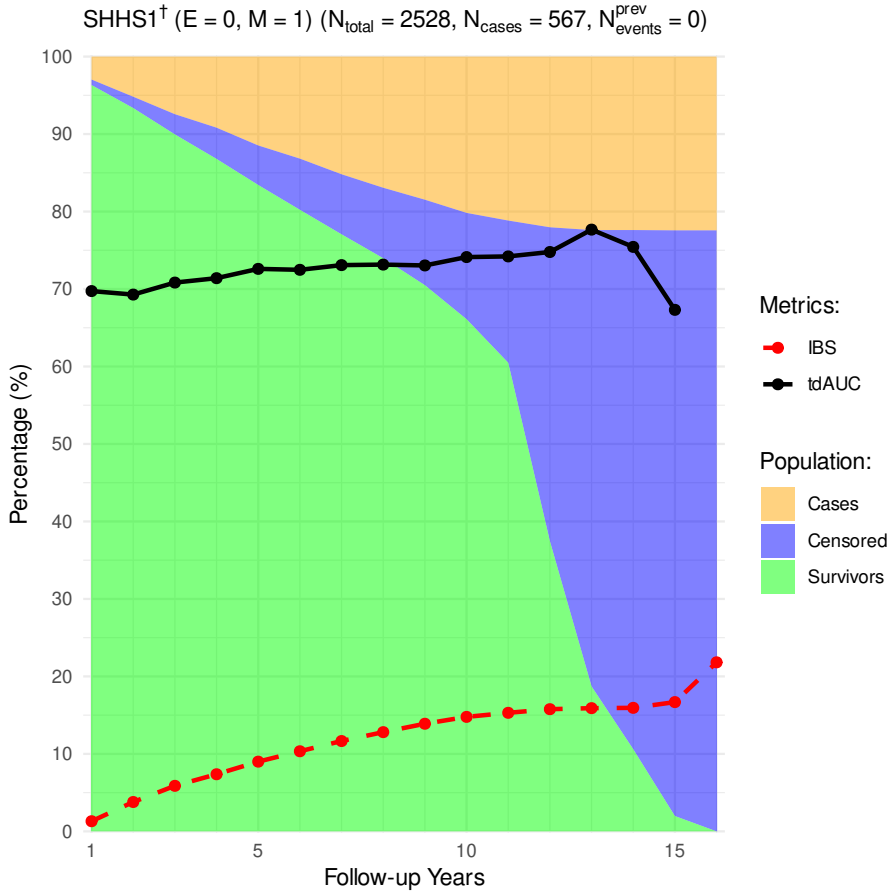
Figure C.1: Cardiovascular outcomes and RSF (including AHI predictor) performance metrics for SHHS1 (E = 0, M = 0).



Notes: The plot shows the distribution of cardiovascular cases, survivors, and censored subjects. Performance is evaluated based on cross-validation (CV), with the time-dependent area under the ROC curve (tdAUROC) and the integrated Brier score (IBS), reported as mean values with standard deviations (SD). A bar over N (e.g., $\overline{N}_{\text{total}}$) denotes the average number of subjects across CV folds.

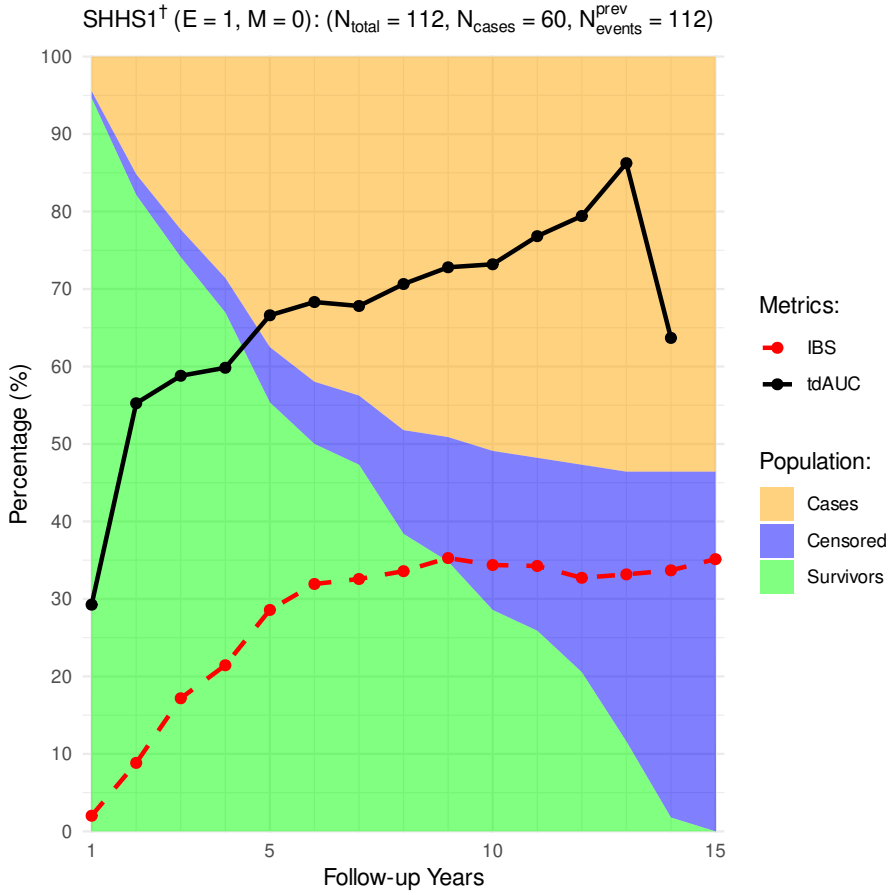
C.4.2 SHHS1 test subjects

Figure C.2: Cardiovascular outcomes and RSF (including AHI predictor) performance metrics for SHHS1†(E = 0, M = 1).



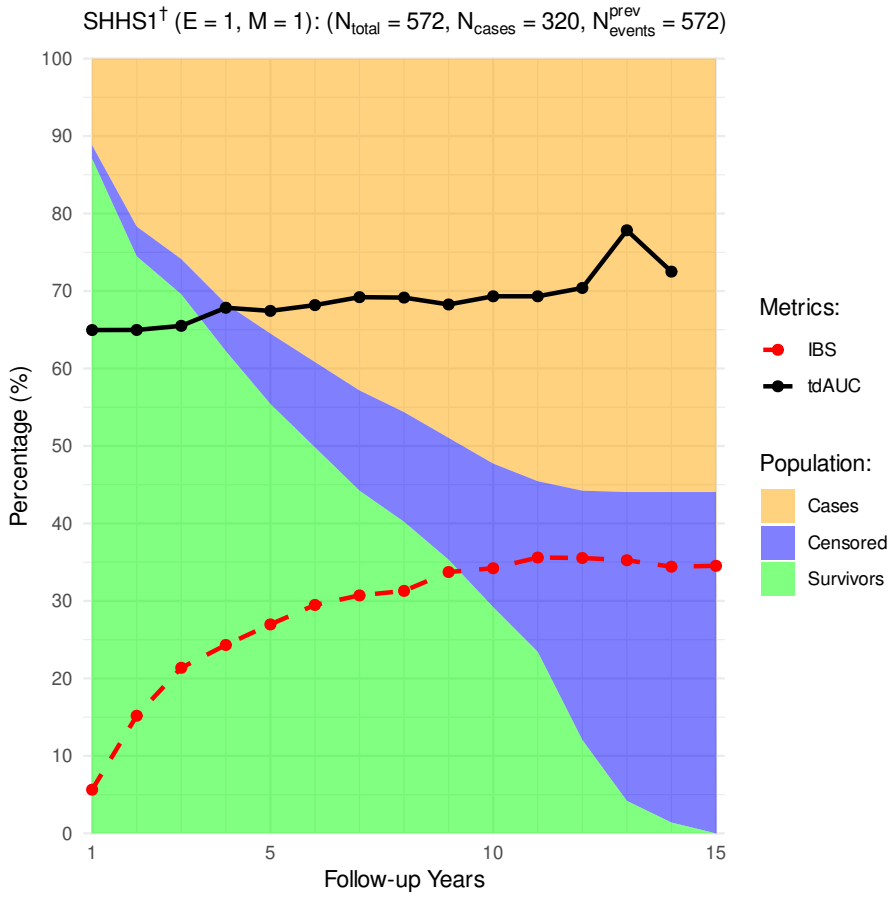
Notes: The plot shows the distribution of cardiovascular cases, survivors, and censored subjects. † denotes evaluation on model-unseen subjects. Performance is evaluated with the time-dependent area under the ROC curve (tdAUROC) and the integrated Brier score (IBS).

Figure C.3: Cardiovascular outcomes and RSF (including AHI predictor) performance metrics for SHHS1†(E = 1, M = 0).



Notes: The plot shows the distribution of cardiovascular cases, survivors, and censored subjects. † denotes evaluation on model-unseen subjects. Performance is evaluated with the time-dependent area under the ROC curve (tdAUROC) and the integrated Brier score (IBS).

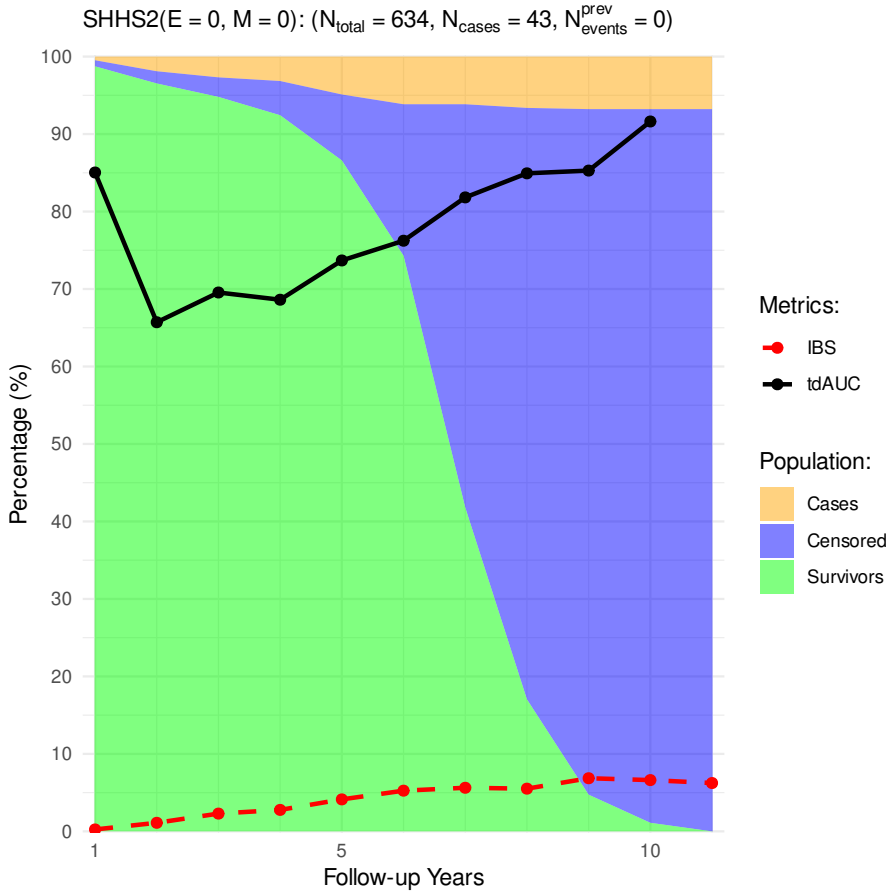
Figure C.4: Cardiovascular outcomes and RSF (including AHI predictor) performance metrics for SHHS1†(E = 1, M = 1).



Notes: The plot shows the distribution of cardiovascular cases, survivors, and censored subjects. † denotes evaluation on model-unseen subjects. Performance is evaluated with the time-dependent area under the ROC curve (tdAUROC) and the integrated Brier score (IBS).

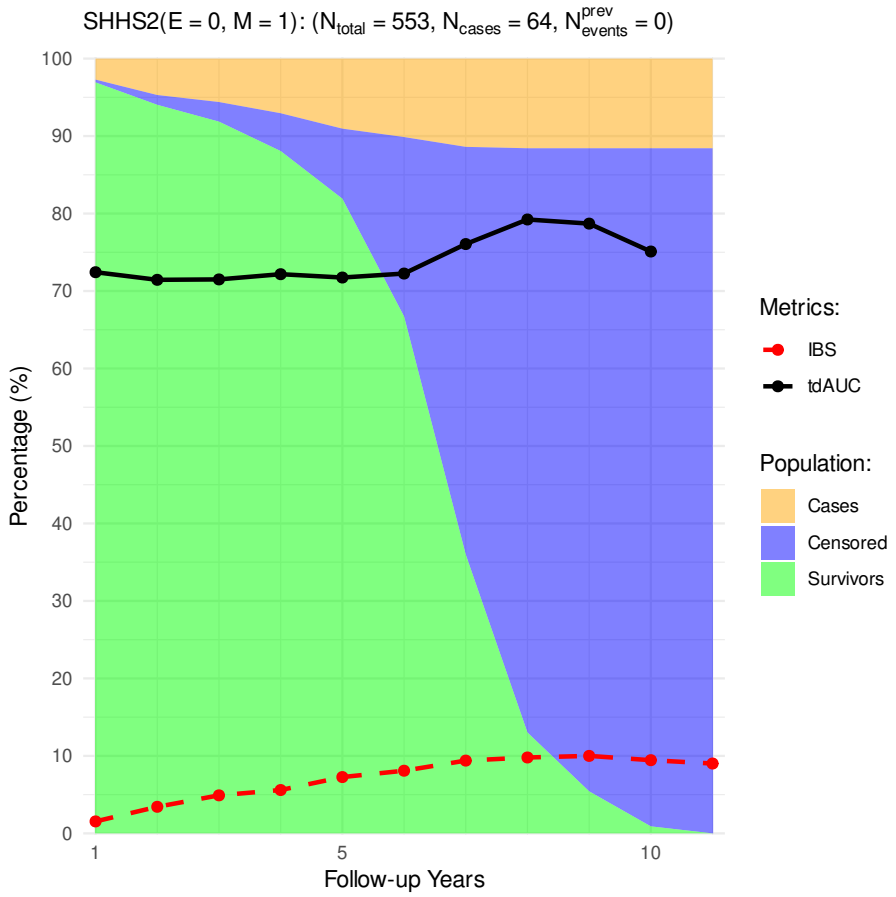
C.4.3 SHHS2 train subjects

Figure C.5: Cardiovascular outcomes and RSF (including AHI predictor) performance metrics for SHHS2 (E = 0, M = 0).



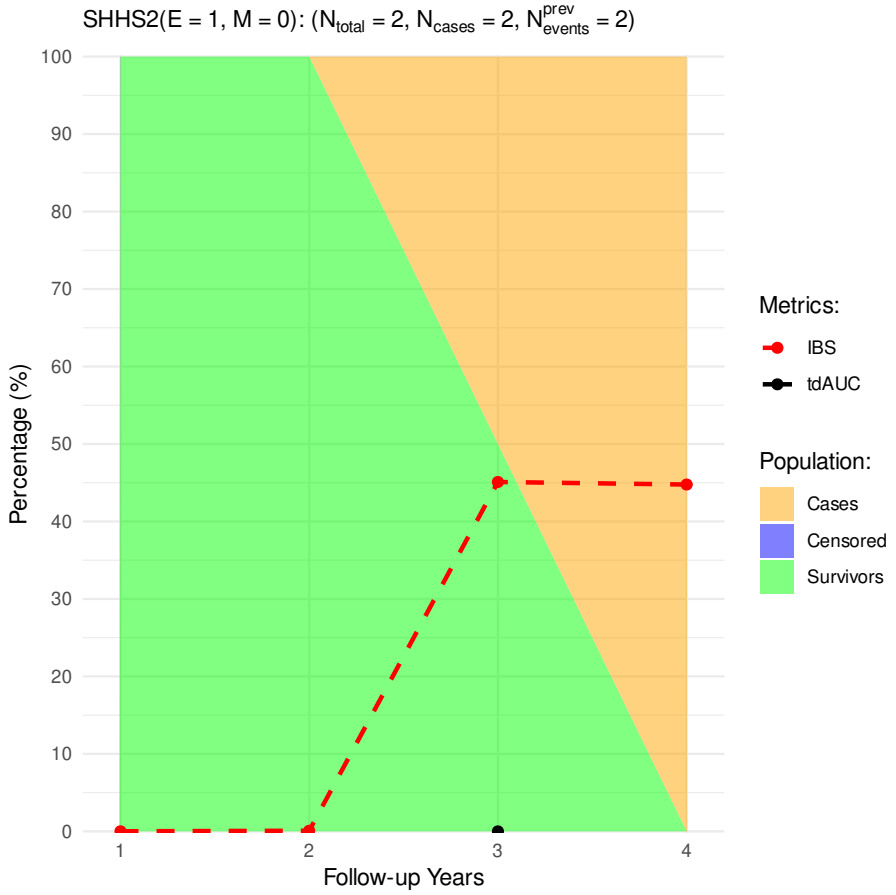
Notes: The plot shows the distribution of cardiovascular cases, survivors, and censored subjects. Performance is evaluated with the time-dependent area under the ROC curve (tdAUROC) and the integrated Brier score (IBS).

Figure C.6: Cardiovascular outcomes and RSF (including AHI predictor) performance metrics for SHHS2 (E = 0, M = 1).



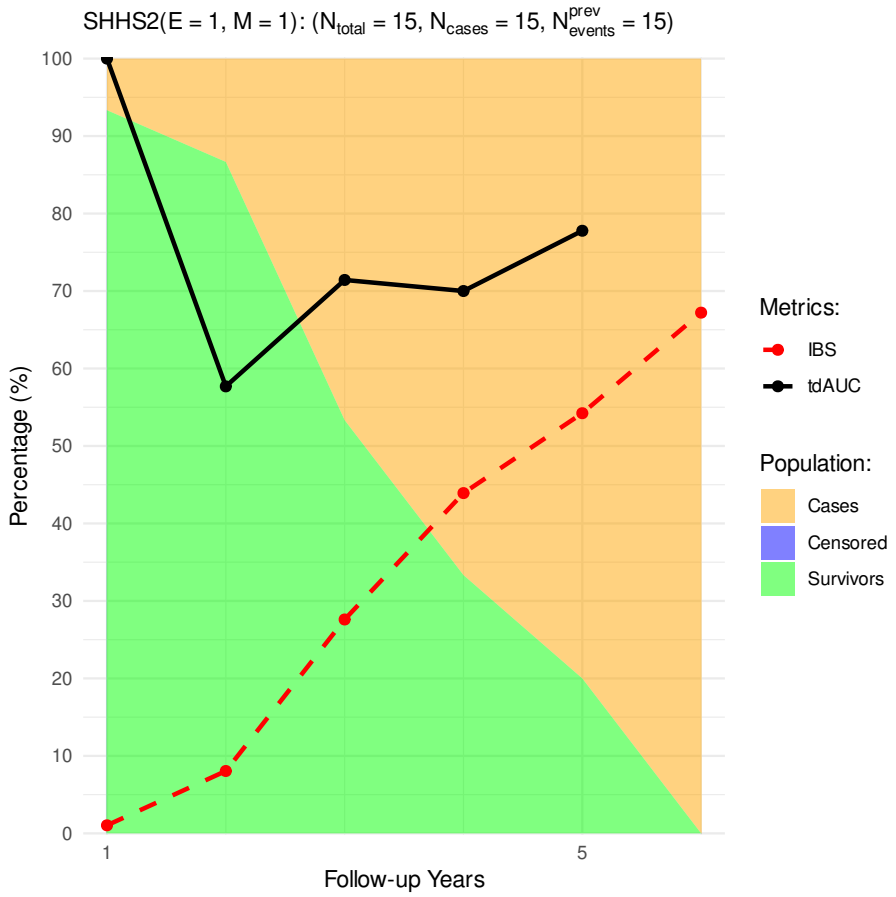
Notes: The plot shows the distribution of cardiovascular cases, survivors, and censored subjects. Performance is evaluated with the time-dependent area under the ROC curve (tdAUROC) and the integrated Brier score (IBS).

Figure C.7: Cardiovascular outcomes and RSF (including AHI predictor) performance metrics for SHHS2 (E = 1, M = 0).



Notes: The plot shows the distribution of cardiovascular cases, survivors, and censored subjects. Performance is evaluated with the time-dependent area under the ROC curve (tdAUROC) and the integrated Brier score (IBS).

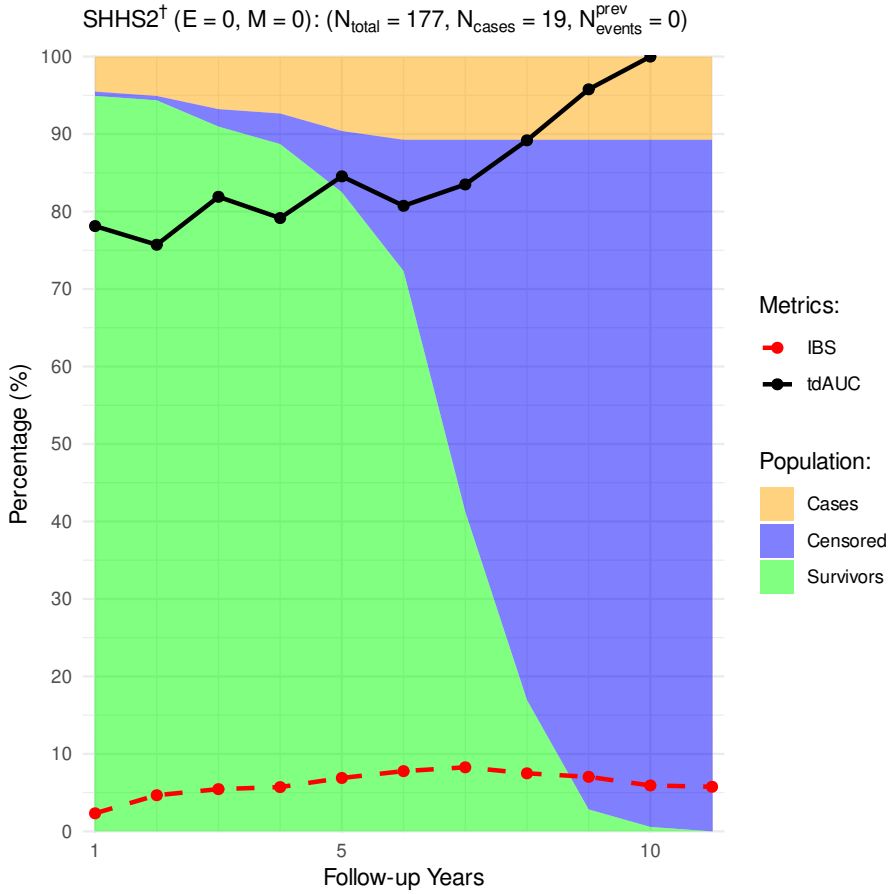
Figure C.8: Cardiovascular outcomes and RSF (including AHI predictor) performance metrics for SHHS2 (E = 1, M = 1).



Notes: The plot shows the distribution of cardiovascular cases, survivors, and censored subjects. Performance is evaluated with the time-dependent area under the ROC curve (tdAUROC) and the integrated Brier score (IBS).

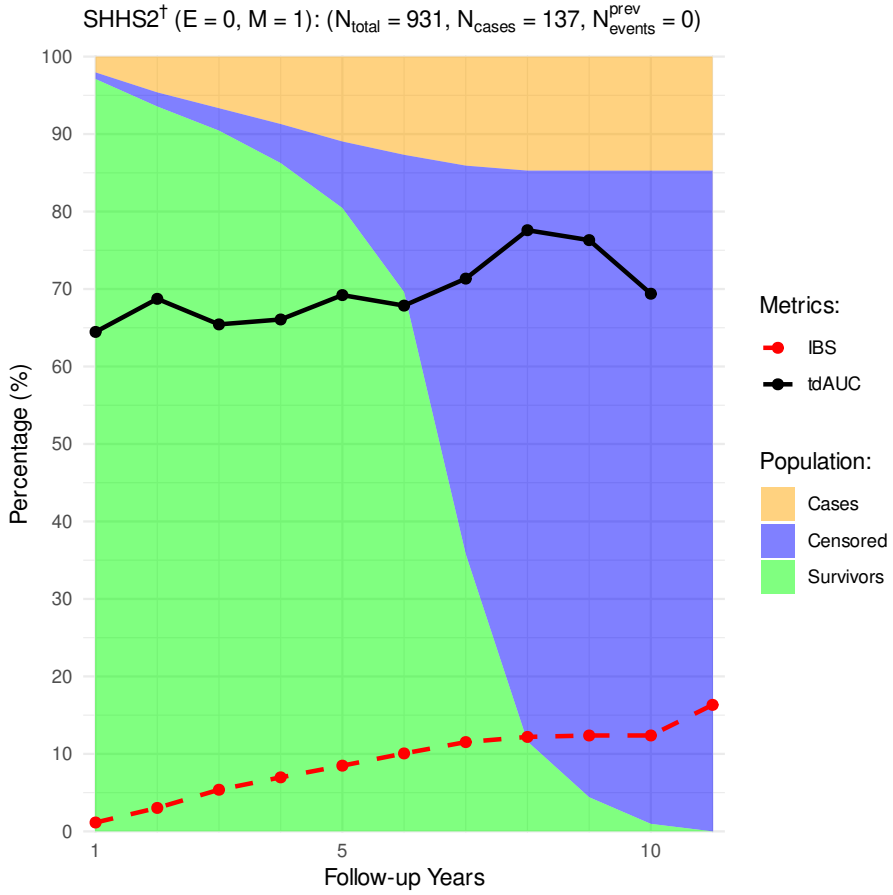
C.4.4 SHHS2 test subjects

Figure C.9: Cardiovascular outcomes and RSF (including AHI predictor) performance metrics for SHHS2†(E = 0, M = 0).



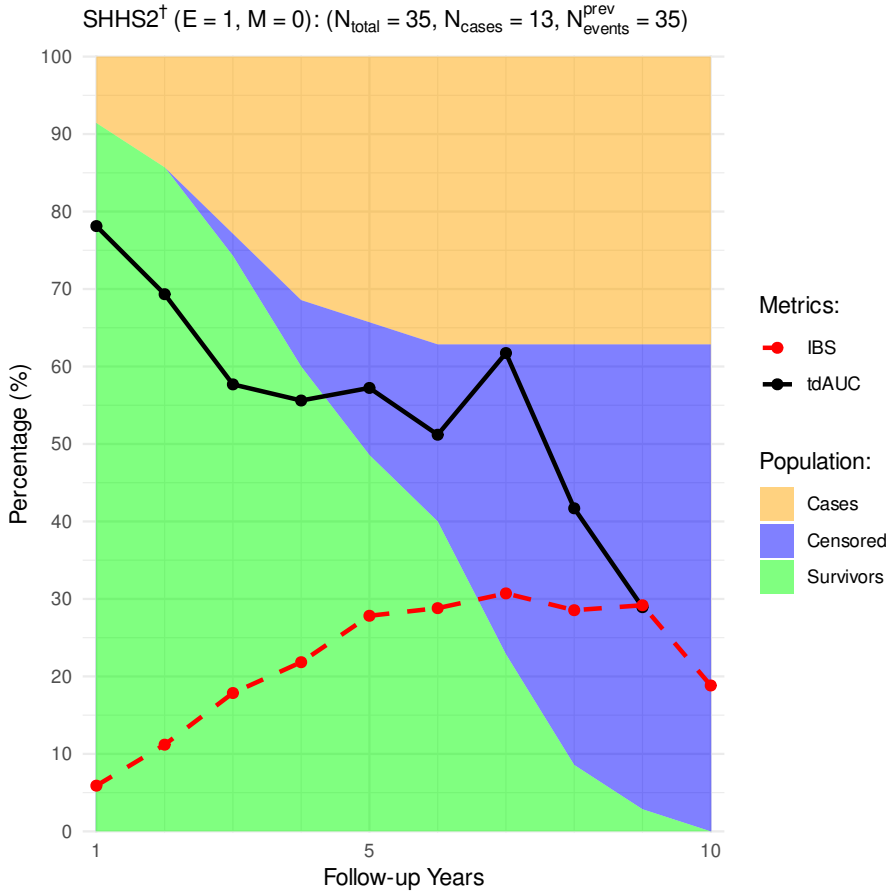
Notes: The plot shows the distribution of cardiovascular cases, survivors, and censored subjects. † denotes evaluation on model-unseen subjects. Performance is evaluated with the time-dependent area under the ROC curve (tdAUROC) and the integrated Brier score (IBS).

Figure C.10: Cardiovascular outcomes and RSF (including AHI predictor) performance metrics for SHHS2†(E = 0, M = 1).



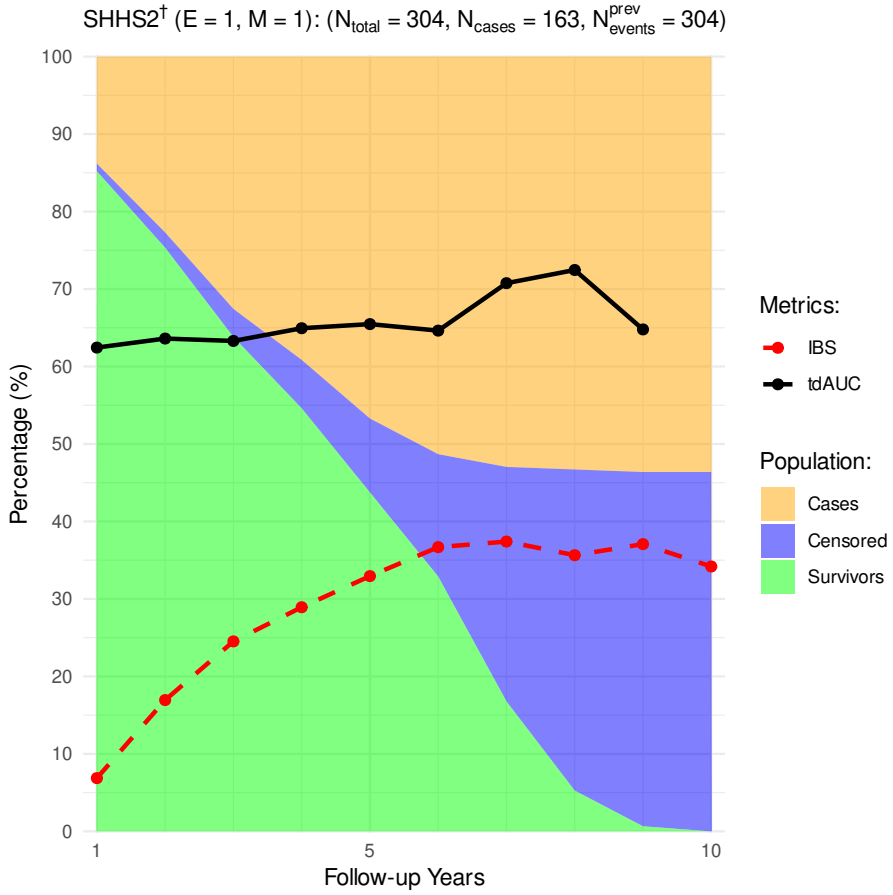
Notes: The plot shows the distribution of cardiovascular cases, survivors, and censored subjects. † denotes evaluation on model-unseen subjects. Performance is evaluated with the time-dependent area under the ROC curve (tdAUROC) and the integrated Brier score (IBS).

Figure C.11: Cardiovascular outcomes and RSF (including AHI predictor) performance metrics for SHHS2†(E = 1, M = 0).



Notes: The plot shows the distribution of cardiovascular cases, survivors, and censored subjects. † denotes evaluation on model-unseen subjects. Performance is evaluated with the time-dependent area under the ROC curve (tdAUROC) and the integrated Brier score (IBS).

Figure C.12: Cardiovascular outcomes and RSF (including AHI predictor) performance metrics for SHHS2†(E = 1, M = 1).

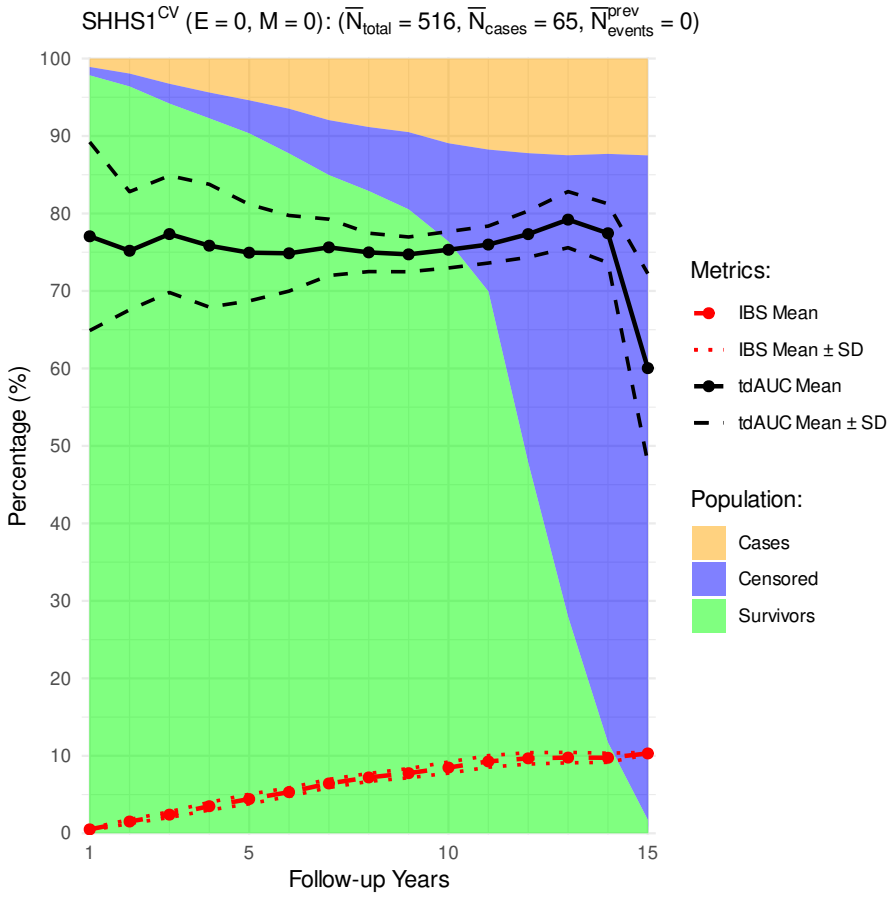


Notes: The plot shows the distribution of cardiovascular cases, survivors, and censored subjects. † denotes evaluation on model-unseen subjects. Performance is evaluated with the time-dependent area under the ROC curve (tdAUROC) and the integrated Brier score (IBS).

C.5 Survival Plots without AHI predictor

C.5.1 Primary study cohort

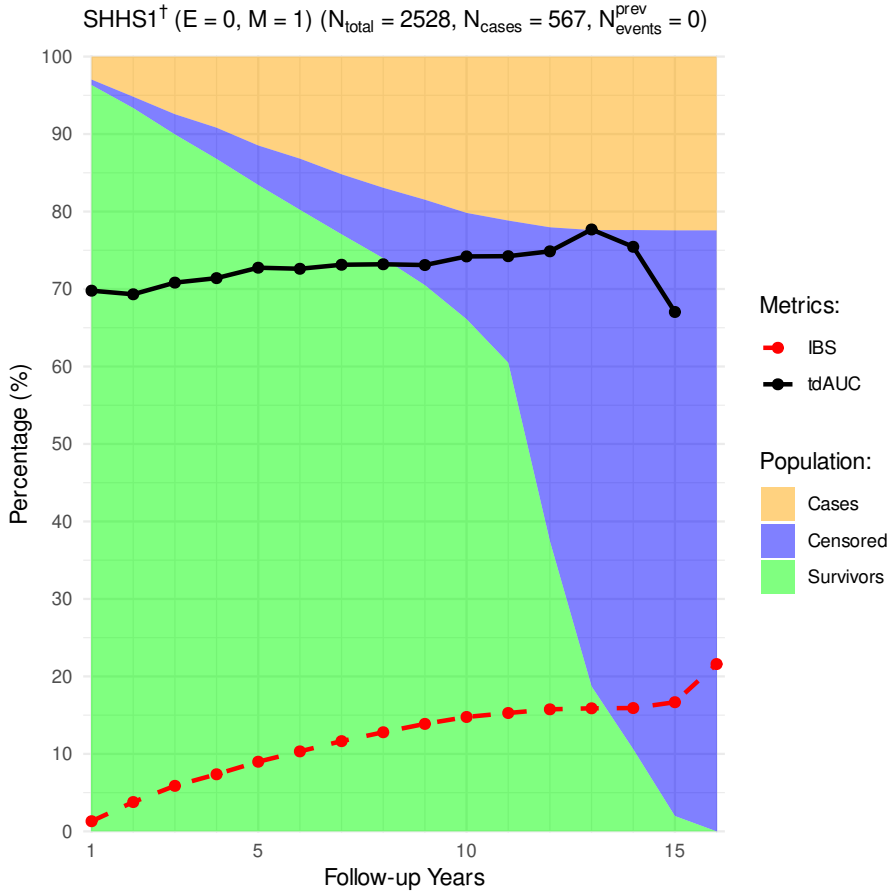
Figure C.13: Cardiovascular outcomes and RSF (excluding AHI predictor) performance metrics for SHHS1 (E = 0, M = 0).



Notes: The plot shows the distribution of cardiovascular cases, survivors, and censored subjects. Performance is evaluated based on cross-validation (CV), with the time-dependent area under the ROC curve (tdAUROC) and the integrated Brier score (IBS), reported as mean values with standard deviations (SD). A bar over N (e.g., $\overline{N}_{\text{total}}$) denotes the average number of subjects across CV folds.

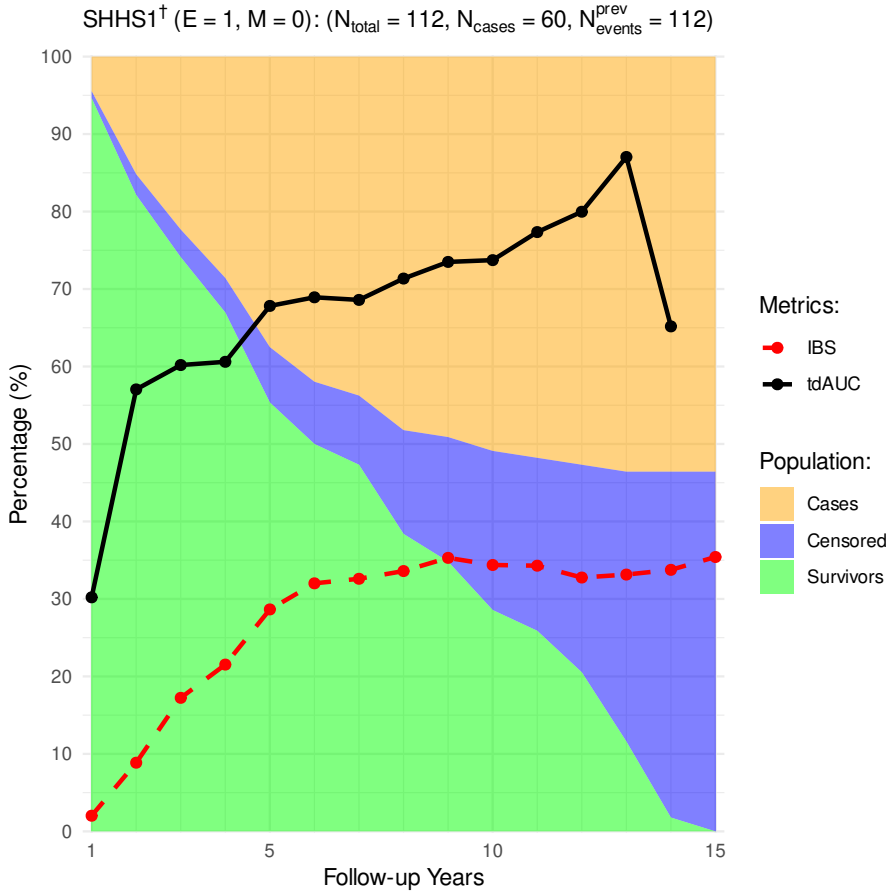
C.5.2 SHHS1 test subjects

Figure C.14: Cardiovascular outcomes and RSF (excluding AHI predictor) performance metrics for SHHS1†(E = 0, M = 1).



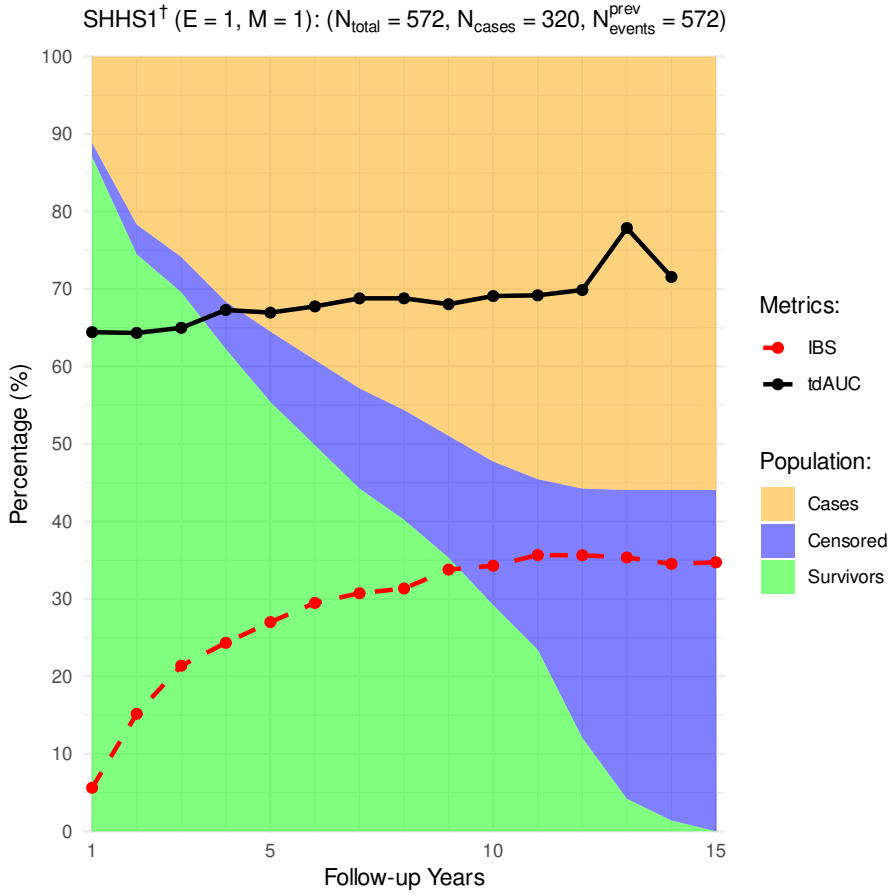
Notes: The plot shows the distribution of cardiovascular cases, survivors, and censored subjects. † denotes evaluation on model-unseen subjects. Performance is evaluated with the time-dependent area under the ROC curve (tdAUROC) and the integrated Brier score (IBS).

Figure C.15: Cardiovascular outcomes and RSF (excluding AHI predictor) performance metrics for SHHS1†(E = 1, M = 0).



Notes: The plot shows the distribution of cardiovascular cases, survivors, and censored subjects. † denotes evaluation on model-unseen subjects. Performance is evaluated with the time-dependent area under the ROC curve (tdAUROC) and the integrated Brier score (IBS).

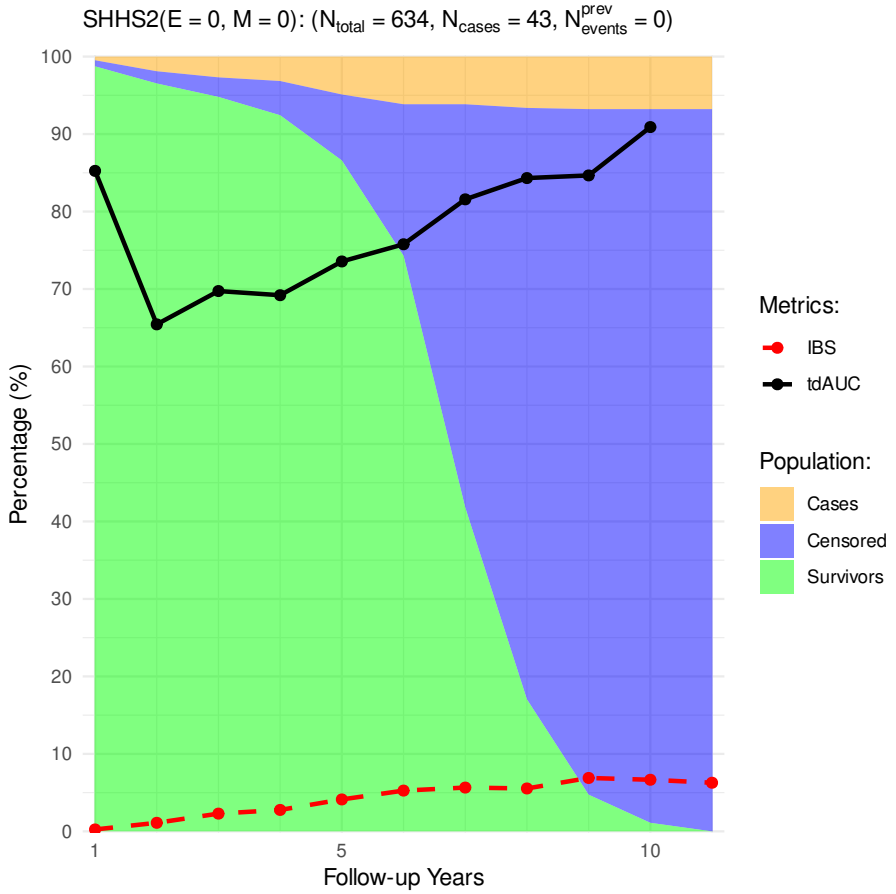
Figure C.16: Cardiovascular outcomes and RSF (excluding AHI predictor) performance metrics for SHHS1†(E = 1, M = 1).



Notes: The plot shows the distribution of cardiovascular cases, survivors, and censored subjects. † denotes evaluation on model-unseen subjects. Performance is evaluated with the time-dependent area under the ROC curve (tdAUROC) and the integrated Brier score (IBS).

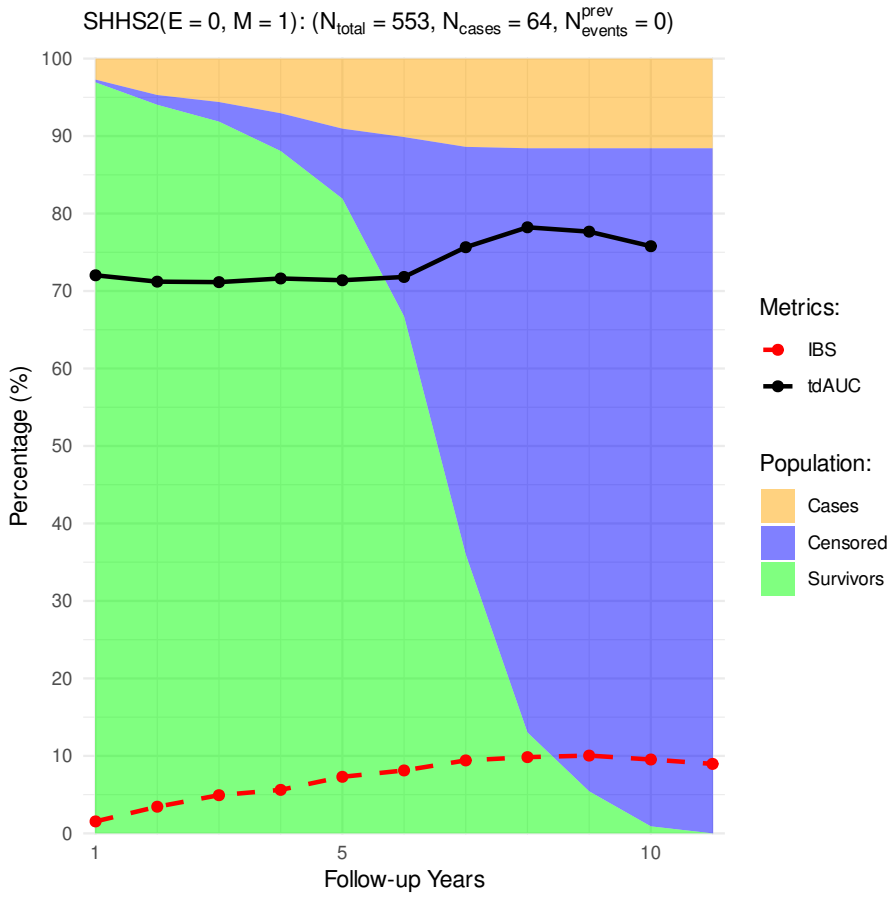
C.5.3 SHHS2 train subjects

Figure C.17: Cardiovascular outcomes and RSF (excluding AHI predictor) performance metrics for SHHS2(E = 0, M = 0).



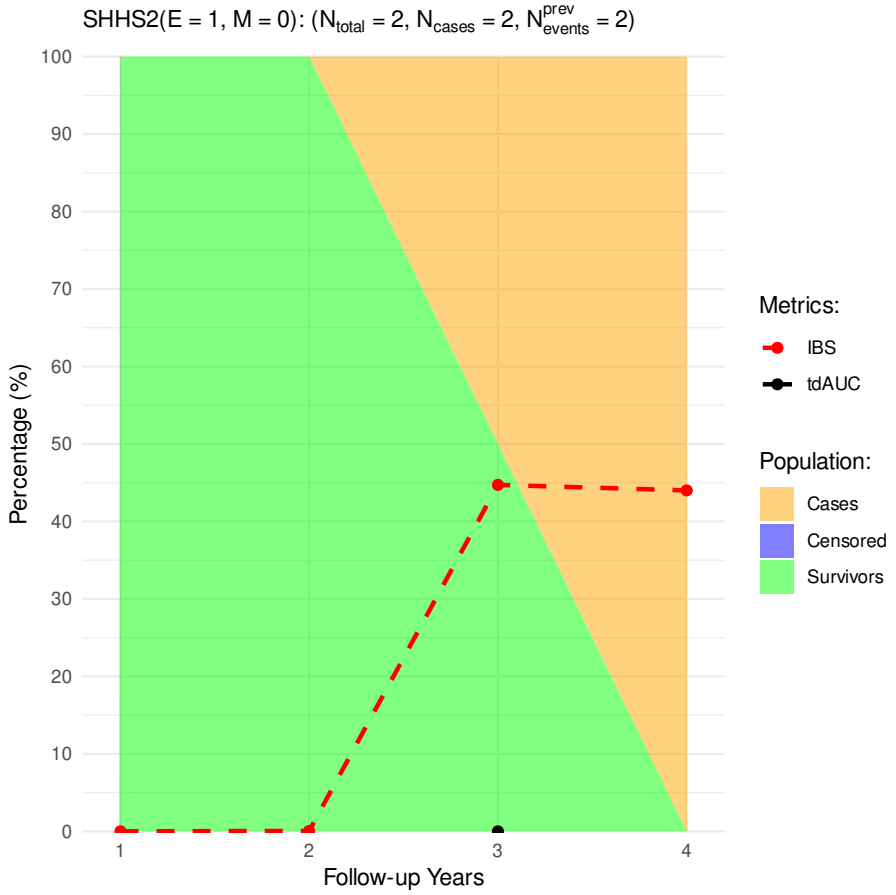
Notes: The plot shows the distribution of cardiovascular cases, survivors, and censored subjects. Performance is evaluated with the time-dependent area under the ROC curve (tdAUROC) and the integrated Brier score (IBS).

Figure C.18: Cardiovascular outcomes and RSF (excluding AHI predictor) performance metrics for SHHS2(E = 0, M = 1).



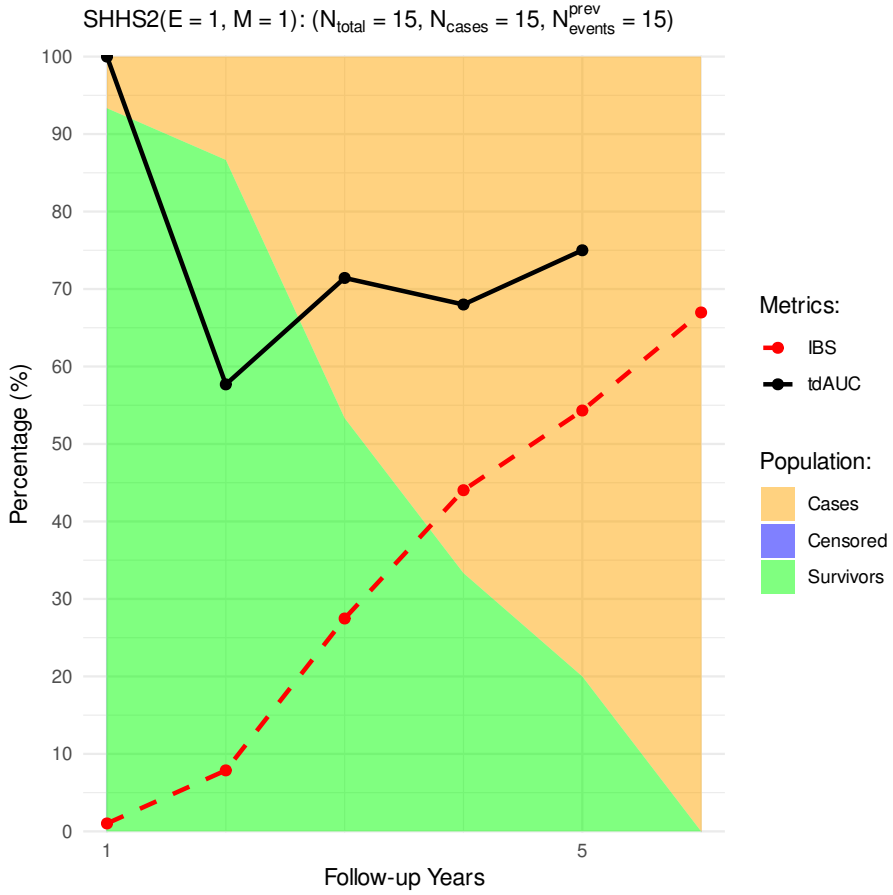
Notes: The plot shows the distribution of cardiovascular cases, survivors, and censored subjects. Performance is evaluated with the time-dependent area under the ROC curve (tdAUROC) and the integrated Brier score (IBS).

Figure C.19: Cardiovascular outcomes and RSF (excluding AHI predictor) performance metrics for SHHS2(E = 1, M = 0).



Notes: The plot shows the distribution of cardiovascular cases, survivors, and censored subjects. Performance is evaluated with the time-dependent area under the ROC curve (tdAUROC) and the integrated Brier score (IBS).

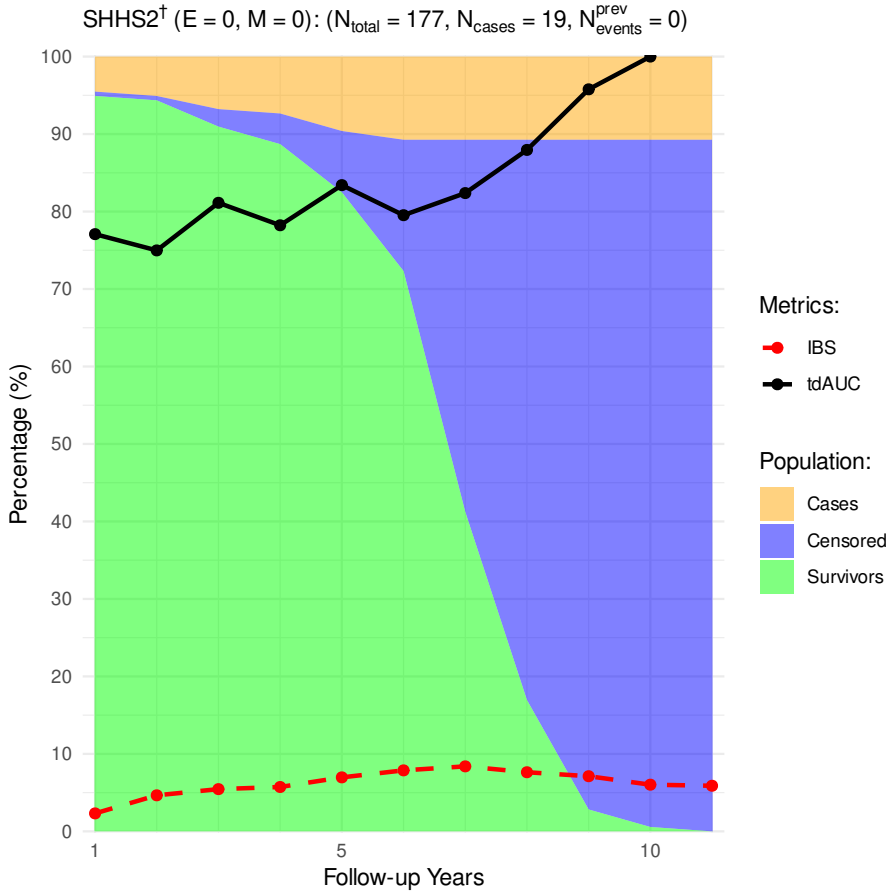
Figure C.20: Cardiovascular outcomes and RSF (excluding AHI predictor) performance metrics for SHHS2(E = 1, M = 1).



Notes: The plot shows the distribution of cardiovascular cases, survivors, and censored subjects. Performance is evaluated with the time-dependent area under the ROC curve (tdAUROC) and the integrated Brier score (IBS).

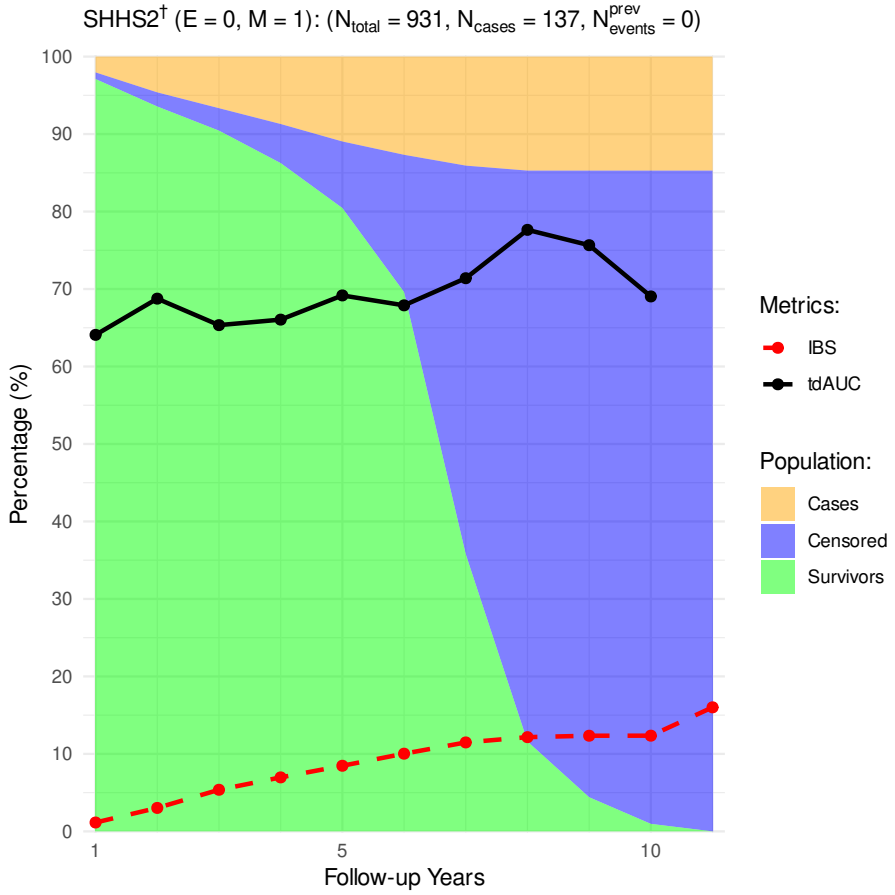
C.5.4 SHHS2 test subjects

Figure C.21: Cardiovascular outcomes and RSF (excluding AHI predictor) performance metrics for SHHS2†(E = 0, M = 0).



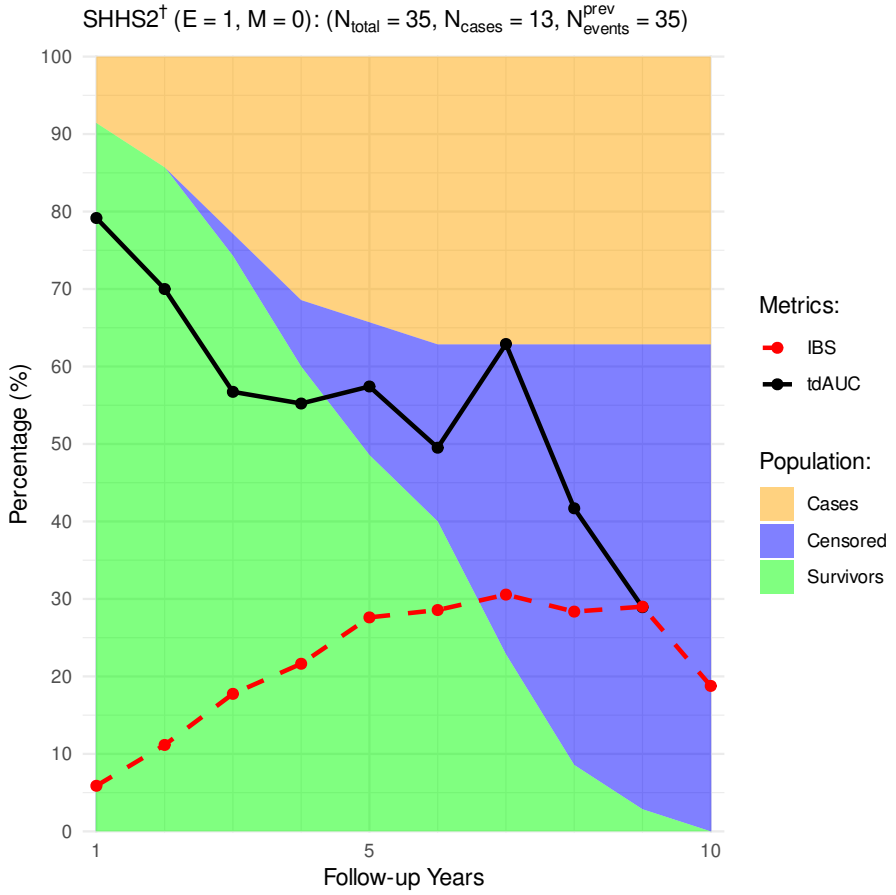
Notes: The plot shows the distribution of cardiovascular cases, survivors, and censored subjects. † denotes evaluation on model-unseen subjects. Performance is evaluated with the time-dependent area under the ROC curve (tdAUROC) and the integrated Brier score (IBS).

Figure C.22: Cardiovascular outcomes and RSF (excluding AHI predictor) performance metrics for SHHS2†(E = 0, M = 1).



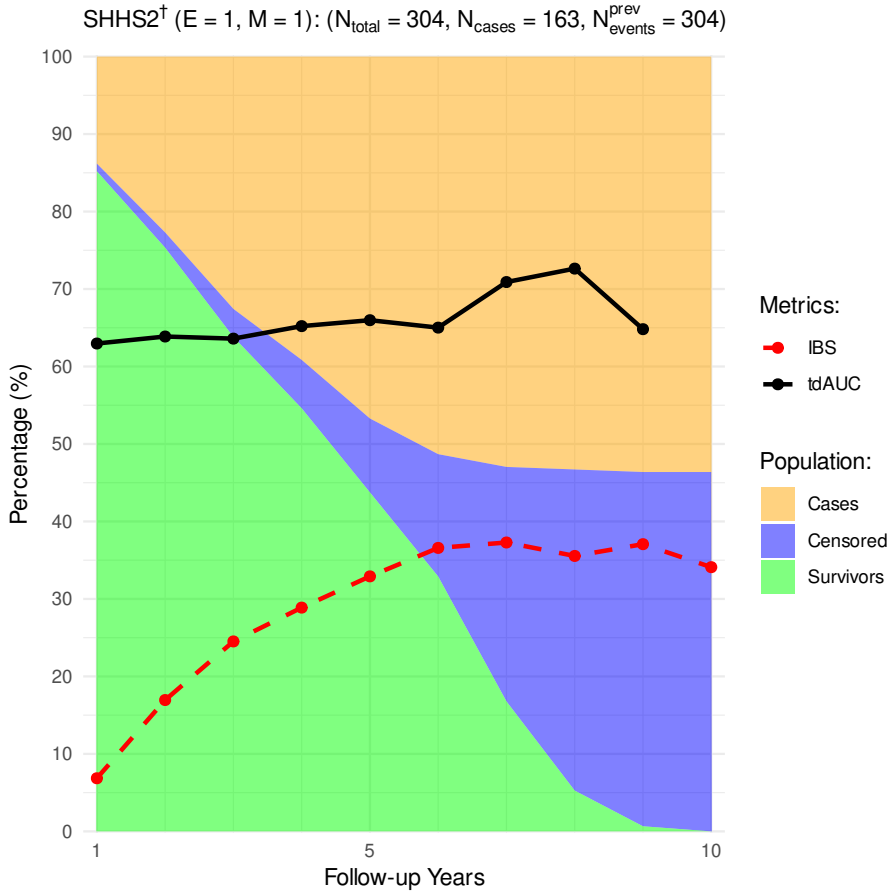
Notes: The plot shows the distribution of cardiovascular cases, survivors, and censored subjects. † denotes evaluation on model-unseen subjects. Performance is evaluated with the time-dependent area under the ROC curve (tdAUROC) and the integrated Brier score (IBS).

Figure C.23: Cardiovascular outcomes and RSF (excluding AHI predictor) performance metrics for SHHS2†(E = 1, M = 0).



Notes: The plot shows the distribution of cardiovascular cases, survivors, and censored subjects. † denotes evaluation on model-unseen subjects. Performance is evaluated with the time-dependent area under the ROC curve (tdAUROC) and the integrated Brier score (IBS).

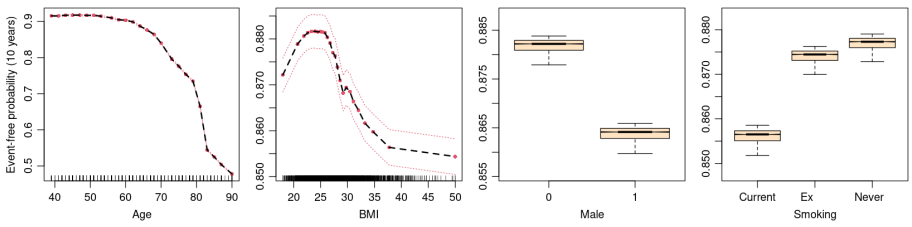
Figure C.24: Cardiovascular outcomes and RSF (excluding AHI predictor) performance metrics for SHHS2†(E = 1, M = 1).



Notes: The plot shows the distribution of cardiovascular cases, survivors, and censored subjects. † denotes evaluation on model-unseen subjects. Performance is evaluated with the time-dependent area under the ROC curve (tdAUROC) and the integrated Brier score (IBS).

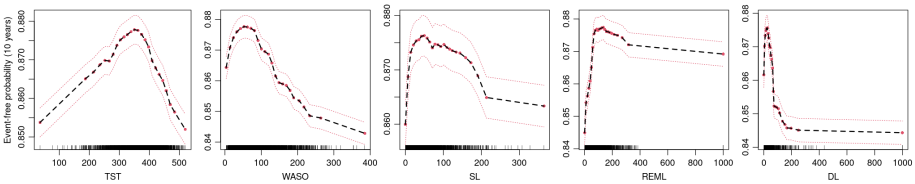
C.6 Partial Effects for RSF without AHI predictor

Figure C.25: Partial effects and their 95% CIs for 10-year cardiovascular event-free probability for the age in years, Body Mass Index (BMI), Apnea-Hypopnea Index (AHI), gender (0 = female, 1 = male), and smoking status, for RSF without AHI predictor.



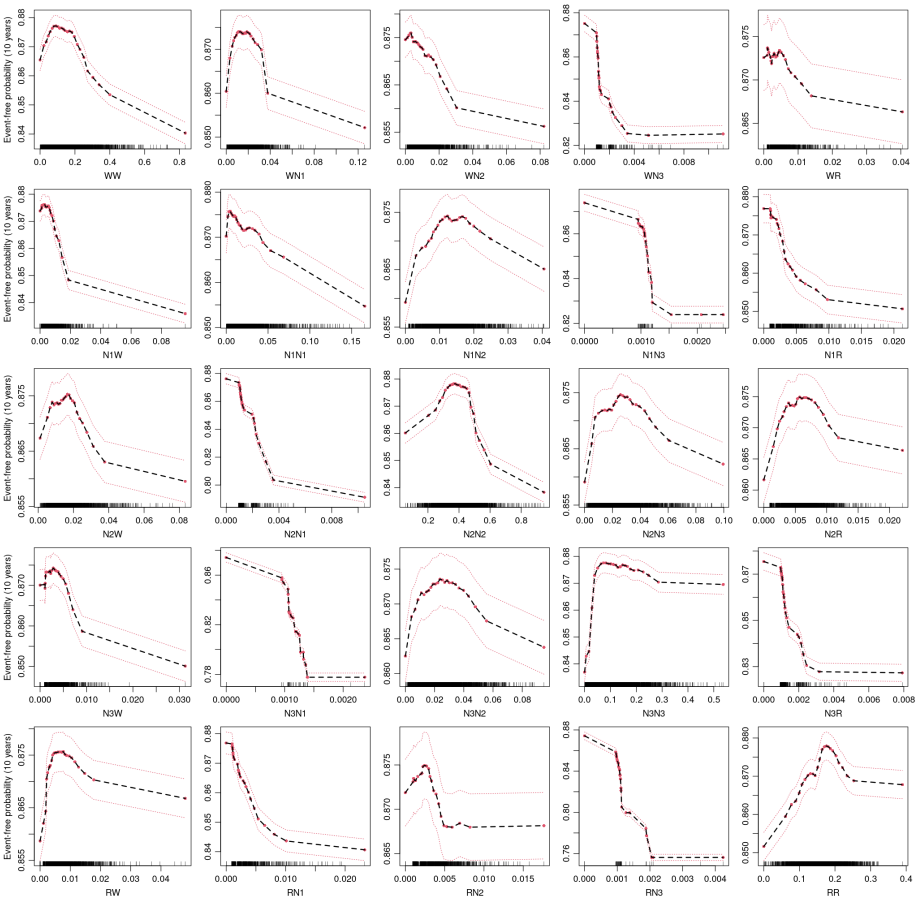
Notes: Data points for continuous predictors are shown as ticks on the x-axis.

Figure C.26: Partial effects and their 95% CIs for 10-year cardiovascular event-free probability for the minutes of Total Sleep Time (TST), Wake After Sleep Onset (WASO), Sleep Latency (SL), REM Latency (REM), and Deep-sleep Latency (DL), for RSF without AHI predictor.



Notes: Data points for continuous predictors are shown as ticks on the x-axis.

Figure C.27: Partial effects and their 95% CIs for 10-year cardiovascular event-free probability for the relative frequencies of transitions between sleep-stage (W, N1, N2, N3, REM), for RSF without AHI predictor.



Notes: Each subplot's x-axis label indicates the transition's direction (e.g., WN1 corresponds to transitions from W to N1). Data points are shown as ticks on the x-axis.