Université Paris Cité (France)

École doctorale Cerveau, Cognition, Comportement (ED 158)
Inserm, F-75019 NeuroDiderot, SleepCmd

In joint doctoral supervision (cotutelle) with Reykjavik University (Iceland)

Department of Computer Science

Characterization of insomnia at home using digital tools and machine learning methods from the combination of objective and subjective criteria

---

By Marie-Ange STEFANOS

PhD Thesis in Neurosciences and Computer Science

Supervised by

Pierre Alexis GEOFFROY and Anna Sigridur ISLIND

Presented and defended publicly on September 14th, 2026

to an examining committee comprised of:

Ludovic Samalin, Prof.	Université Clermont Auvergne, France	Referee
Danny Eckert, Prof.	Flinders University, Australia	Referee
Bastien Lechat, Senior Research Fellow	Flinders University, Australia	Examiner
Jean-Baptiste Masson, Prof.	Institut Pasteur, France	Invited member
Chantal Henry, Prof.	Université Paris Cité, France	Invited member
Pierre Alexis Geoffroy, Prof.	Université Paris Cité, France	Co-supervisor
Anna Sigridur Islind, Prof.	Reykjavik University, Iceland	Co-supervisor
Guillaume de Laboulaye, Head of Machine Learning	Withings, France	Invited member

© 2026 Marie-Ange Stefanos

Characterization of insomnia at home using digital tools and machine learning methods from the combination of objective and subjective criteria

ISBN 978-9935-577-19-1 (print)
ISBN 978-9935-577-20-7 (electronic)

ORCID: 0009-0001-8903-1537

Abstract

Title:

Characterization of insomnia at home using digital tools and machine learning methods from the combination of objective and subjective criteria

Keywords:

Insomnia ; Sleep assessment ; Digital health technologies ; Wearable and nearable sensors ; Smart devices ; Actigraphy ; Machine learning ; Unsupervised clustering ; Longitudinal at-home monitoring ; Subjective and objective sleep measures

Abstract:

Insomnia is one of the most common sleep disorders, degrading quality of life and raising the risk of accidents and of cardiovascular and metabolic disease. Its assessment is split between objective measures such as sleep latency, wake after sleep onset, and total sleep time, obtained from polysomnography (PSG) or actigraphy, and the subjective complaint captured by questionnaires such as the Insomnia Severity Index (ISI). The two correlate poorly: many who report disturbed sleep meet no objective criteria, and some who sleep normally by self-report do. Insomnia is diagnosed from the subjective complaint, actigraphy tracking sleep-wake rhythms, with PSG reserved for detecting sleep abnormalities and comorbidities; this reference examination is costly, depends on scarce expert scorers, and covers one or two laboratory nights, so it reflects neither the home environment nor the night-to-night variability that characterizes the disorder.

Connected sleep-tracking devices used at home offer a complementary route. The under-mattress Sleep Analyzer and the wrist-worn ScanWatch record activity and cardiorespiratory markers continuously over long periods and, when paired with questionnaires, support an assessment that integrates objective and subjective criteria. This thesis investigates how such tools, combined with machine learning, can characterize insomnia in ecological conditions and search for homogeneous subgroups of the disorder.

The work comprises four contributions, each an independent publication. The first is a systematic review of digital tools for insomnia detection and characterization, which maps the signals, datasets, models, and validation practices in the literature and identifies recurrent weaknesses, including small datasets, poorly documented reference labels, and subject-level data leakage that inflates reported performance. The second evaluates the Sleep Analyzer sleep-stage algorithm against PSG in an at-home study. The third designs and runs the SLEEPI protocol, an ecological longitudinal study pairing objective device recordings with repeated questionnaires, among them the ISI, PHQ-9, and GAD-7, with labels from DSM-5 criteria. The fourth is a prospective analysis of the longitudinal device data that searches for insomnia subgroups in the objective sleep and activity features. Under unsupervised clustering these features form a single continuum rather than discrete subtypes, and the objective signal that tracks self-reported severity is weak, so the analysis reframes the objective phenotype as graded and points to within-person night-to-night variability as a dimension that longitudinal home monitoring can measure and that warrants dedicated future investigation. The thesis thus argues that longitudinal, at-home, multimodal monitoring is a practical path toward a more individualized characterization of insomnia than current assessment methods allow.

Résumé

Titre :

Caractérisation de l'insomnie à domicile à l'aide de solutions numériques et de méthodes de machine learning à partir du croisement de critères objectifs et subjectifs

Mots-clés :

Insomnie ; Évaluation du sommeil ; Technologies numériques de santé ; Capteurs wearables et nearables ; Objets connectés ; Actigraphie ; Apprentissage automatique ; Clustering non supervisé ; Suivi longitudinal à domicile ; Mesures subjectives et objectives du sommeil

Résumé :

L'insomnie est l'un des troubles du sommeil les plus fréquents ; elle dégrade la qualité de vie et augmente le risque d'accidents ainsi que de maladies cardiovasculaires et métaboliques. Son évaluation se base sur des mesures objectives telles que la latence d'endormissement, le temps d'éveil après endormissement et le temps de sommeil total, obtenues par polysomnographie (PSG) ou actigraphie, et la plainte subjective recueillie par des questionnaires tels que l'indice de sévérité de l'insomnie (ISI). Les deux sont peu corrélés : de nombreuses personnes qui souffrent d'un sommeil perturbé ne remplissent pas les critères objectifs, tandis que certaines sans plainte de sommeil peuvent les remplir. L'insomnie est diagnostiquée à partir de la plainte subjective, l'actigraphie offrant un suivi des rythmes veille-sommeil, la PSG étant réservée à la détection des anomalies du sommeil et des comorbidités. Cet examen de référence est coûteux, repose sur une annotation manuelle assurée par des spécialistes peu nombreux et se limite à une ou deux nuits de laboratoire. Il ne reflète donc pas l'environnement du domicile ni la variabilité d'une nuit à l'autre qui caractérise le trouble.

Les dispositifs connectés de suivi du sommeil utilisés à domicile offrent un potentiel outil complémentaire. Le Sleep Analyzer, placé sous le matelas, et la ScanWatch, portée au poignet, enregistrent en continu et sur de longues périodes l'activité et des marqueurs cardiorespiratoires et, associés à des questionnaires, permettent une évaluation qui intègre des critères objectifs et subjectifs. Cette thèse étudie comment de tels outils, combinés à l'apprentissage automatique, peuvent caractériser l'insomnie en conditions écologiques et rechercher des sous-groupes homogènes du trouble.

Le travail s'articule autour de quatre contributions, chacune sous forme d'une publication indépendante. La première est une revue systématique des outils numériques pour la détection et la caractérisation de l'insomnie, qui recense les signaux, les jeux de données, les modèles et les méthodes de validation de la littérature et met en évidence des axes d'amélioration, notamment des jeux de données de petite taille, des labels peu documentés et une fuite de données au niveau des sujets qui gonfle artificiellement les performances rapportées. La deuxième évalue l'algorithme d'identification des stades de sommeil du Sleep Analyzer face à la PSG dans une étude à domicile. La troisième décrit la conception et la réalisation du protocole SLEEPI, une étude de collecte de données longitudinale et écologique associant des enregistrements objectifs issus des dispositifs à des questionnaires subjectifs répétés, parmi lesquels l'ISI, le PHQ-9 et le GAD-7, avec des labels dérivés des critères du DSM-5. La quatrième est une analyse prospective des données longitudinales des dispositifs qui recherche des sous-groupes d'insomnie dans les caractéristiques objectives de sommeil et d'activité. Sous clustering non supervisé, ces caractéristiques forment un continuum plutôt que des sous-types distincts, et les données objectives et la sévérité d'insomnie subjective sont peu corrélées. L'analyse redéfinit donc le phénotype objectif comme graduel et désigne la variabilité intra-individuelle d'une nuit à l'autre comme une dimension que le suivi longitudinal à domicile permet de mesurer et qui mérite une investigation future dédiée. L'extension de l'analyse de ces agrégats par sujet vers l'ensemble des séries temporelles de sommeil est laissée à de futurs travaux. La thèse soutient ainsi qu'un suivi longitudinal, à domicile et multimodal constitue une voie concrète vers une caractérisation plus individualisée de l'insomnie que ne le permettent les méthodes d'évaluation actuelles.

Résumé substantiel en français

Contexte

L'insomnie est l'un des troubles du sommeil les plus fréquents ; elle dégrade la qualité de vie et augmente le risque d'accidents ainsi que de maladies cardiovasculaires et métaboliques. Son évaluation repose sur deux natures de données. D'un côté, des mesures objectives telles que la latence d'endormissement, le temps d'éveil après endormissement et le temps de sommeil total, obtenues par polysomnographie (PSG) ou par actigraphie ; de l'autre, la plainte subjective recueillie par des questionnaires tels que l'indice de sévérité de l'insomnie (ISI). Ces deux natures de données sont peu corrélées : de nombreuses personnes qui souffrent d'un sommeil perturbé ne remplissent pas les critères objectifs, tandis que certaines sans plainte de sommeil les remplissent. En pratique, l'insomnie est diagnostiquée à partir de la plainte subjective, l'actigraphie offrant un suivi des rythmes veille-sommeil, la PSG étant réservée à la détection des anomalies du sommeil et des comorbidités. Cet examen de référence est coûteux, repose sur une annotation manuelle des signaux assurée par des spécialistes peu nombreux et se limite à une ou deux nuits en laboratoire. Il ne reflète donc ni l'environnement du domicile ni la variabilité d'une nuit à l'autre qui caractérise le trouble.

Objectifs

Les dispositifs connectés de suivi du sommeil utilisés à domicile offrent un potentiel outil complémentaire. Le Sleep Analyzer, placé sous le matelas, et la ScanWatch, portée au poignet, enregistrent en continu et sur de longues périodes l'activité et des marqueurs cardiorespiratoires. Associés à des questionnaires, ils permettent une évaluation qui croise des critères objectifs et subjectifs. Cette thèse étudie comment de tels outils, combinés à l'apprentissage automatique, peuvent caractériser l'insomnie en conditions écologiques et rechercher des sous-groupes homogènes du trouble. Le travail s'articule autour de quatre contributions, chacune préparée sous la forme d'une publication indépendante : une revue systématique de la littérature, l'évaluation d'un dispositif de mesure du sommeil, la conception d'un protocole de collecte longitudinale de données à domicile et à grande échelle, et l'analyse des données recueillies.

Revue systématique des outils numériques

La première contribution est une revue systématique des outils numériques pour la détection et la caractérisation de l'insomnie. À partir de quatre plateformes bibliographiques,

4 184 références uniques ont été examinées et 49 études retenues, la recherche s'achevant en février 2024. Ces 49 études se répartissent en quatre types de sorties qui ne sont pas directement comparables : la classification binaire dans 17 études, la classification multi-classes ou multi-tâches dans 10, le phénotypage par clustering dans 13, et l'analyse statistique ou de traitement du signal dans 9. La revue met en évidence des axes d'amélioration. 65 % des études s'appuient sur des jeux de données propriétaires et aucun dataset ne fait consensus. 59 % reposent sur une durée d'enregistrement d'au plus 48 heures, ce qui ne permet pas d'évaluer la variabilité d'une nuit à l'autre. Les données subjectives ne sont collectées et exploitées que rarement. Le *feature engineering* domine : seules 3 des 49 études utilisent un apprentissage profond de bout en bout, toutes sur l'électroencéphalogramme. L'hétérogénéité des types de sorties, des définitions de l'insomnie et des jeux de données empêche toute comparaison des performances entre études, d'autant que 11 des 49 études proviennent de trois groupes de recherche seulement, si bien que le nombre apparent d'études surestime le nombre de sources indépendantes. Le suivi multi-nuits et l'intégration conjointe des données subjectives et objectives sont identifiés comme des priorités pour les travaux futurs.

Évaluation de l'algorithme du Sleep Analyzer

La deuxième contribution évalue l'algorithme d'identification des stades de sommeil du Sleep Analyzer face à la PSG. L'étude, prospective et menée à domicile, porte sur 117 adultes sains (69 femmes ; âge moyen 39,9 ans, écart-type 11,4) enregistrés simultanément par PSG en conditions de vie réelle, sans restriction sur la routine du participant (médicaments, alcool, caféine, etc.). La distinction veille–sommeil atteint une *accuracy* de 87 % (IC à 95 % 86–87 %), avec une sensibilité élevée pour le sommeil (93 %) et modérée pour l'éveil (73 %). L'identification des stades atteint une *accuracy* moyenne de 63 % (IC à 95 % 62–65 %) et un coefficient κ de Cohen de 0,49, la plupart des erreurs concernant l'identification correcte des phases de sommeil léger. Le dispositif surestime le temps de sommeil total de 20 minutes et le sommeil léger, et sous-estime le sommeil paradoxal et le sommeil profond. Ses performances restent stables selon l'âge, le sexe, l'indice de masse corporelle (BMI), la qualité de sommeil auto-évaluée, l'épaisseur du matelas et la présence d'un partenaire de lit. Les erreurs d'identification de stades concernent une confusion avec le sommeil léger et reflètent les désaccords entre experts, qui s'accordent le moins sur ce stade, de sorte que l'*accuracy* est bornée par la variabilité inter-annotateurs des labels de référence, rendant la référence elle-même intrinsèquement imparfaite. L'évaluation montre néanmoins qu'une mesure des stades de sommeil à domicile sans contact direct avec l'utilisateur est possible avec une *accuracy* utilisable, tout en soulignant le besoin d'une méthodologie d'évaluation standardisée, en particulier pour les populations au sommeil perturbé. La comparaison des performances entre études reste difficile, les critères d'exclusion des sujets et la définition de la fenêtre temporelle d'analyse variant fortement d'une étude à l'autre.

Protocole SLEEPI

La troisième contribution décrit la conception et la réalisation du protocole SLEEPI, une étude clinique décentralisée de collecte de données à grande échelle. Sur une période

de 26 semaines et en conditions écologiques, le protocole recueille deux flux de données synchronisés : des mesures physiologiques objectives issues de dispositifs Withings (le Sleep Analyzer sous le matelas ainsi que les montres ScanWatch, enregistrant l'architecture du sommeil, l'activité, la température, la fréquence cardiaque, sa variabilité, la saturation en oxygène et le ronflement) et des questionnaires subjectifs délivrés dans l'application (critères d'insomnie du DSM-5, ISI, PHQ-9, GAD-7, ainsi que des données démographiques, cliniques et environnementales). Les rapports des participants fournissent les labels : les critères du DSM-5 définissent une classification binaire insomnie/bon dormeur et l'ISI une tâche de régression de la sévérité. Le dimensionnement du nombre de sujets à recruter suit une stratégie en deux temps : une simulation détermine d'abord l'effectif minimal requis pour l'objectif d'apprentissage automatique, puis un modèle prospectif d'attrition calcule le nombre de contacts initiaux nécessaires pour l'atteindre. L'ensemble du parcours du participant, du recrutement au retrait, est automatisé et mené avec une intervention humaine réduite par rapport à une étude clinique traditionnelle. Le protocole a permis le suivi continu de plus de 10 000 individus, constituant le jeu de données longitudinal sur l'insomnie le plus vaste et le plus divers à ce jour à notre connaissance, avec des taux de recrutement supérieurs aux prévisions. La population, issue d'utilisateurs Withings, peut différer de la population générale, ce qui limite la généralisation des résultats. Le recours à des dispositifs grand public en conditions réelles privilégie l'échelle et la durée de collecte au prix d'une précision plus faible et d'une adhérence variable qui appellent un prétraitement substantiel des données. Le protocole offre ainsi un modèle reproductible et transférable à d'autres domaines de santé, comme le suivi de patients atteints de maladies cardiovasculaires, la progression des maladies neurodégénératives, la gestion de la douleur chronique, ou encore la santé mentale.

Analyses préliminaires des données de la cohorte SLEEPI

La quatrième contribution est une analyse prospective des données longitudinales, qui recherche des sous-groupes d'insomnie dans les caractéristiques objectives de sommeil et d'activité. Les analyses subjectives portent sur une cohorte analysable de 10 878 participants et les analyses objectives sur une cohorte filtrée de 8 738 sujets disposant de données exploitables. Les scores subjectifs sont cohérents entre eux et gradués selon la sévérité : l'ISI est corrélé au PHQ-9 à $\rho = 0{,}61$ et au GAD-7 à $\rho = 0{,}51$ (corrélation de Spearman), et sa médiane va de 4,5 chez les bons dormeurs à 12,3 dans l'insomnie secondaire à travers les classes du DSM-5. La comorbidité croît avec l'insomnie, la dépression concernant 23 % des sujets et l'anxiété 14 % parmi ceux souffrant d'insomnie secondaire, contre 5 % et 3 % chez les bons dormeurs. Les variables objectives décrivent un espace de grande dimension : le temps de sommeil total médian est de 7,32 heures, l'efficacité du sommeil de 0,97 et la stabilité inter-journalière de 0,33. Une analyse en composantes principales confirme l'absence d'axe dominant, ses deux premières composantes n'expliquant que 39,3 % de la variance. La variabilité d'une nuit à l'autre est importante, avec un écart-type intra-individuel de 1,12 heure pour le temps de sommeil total, tandis que l'ISI reflète surtout une caractéristique stable de la personne : sa corrélation intra-classe de 0,75 indique qu'environ un quart seulement de sa variance provient des fluctuations d'une administration à l'autre. Aucun phénotype discret n'émerge : la meilleure partition non dégénérée n'atteint qu'une silhouette de 0,21 avec un accord quasi nul avec les labels cliniques (ISI, DSM-5), et un espace

latent d'autoencodeur sur lequel a été réalisée une tentative de clustering via HDBSCAN est entièrement classé comme bruit par un pipeline testé sur des données contrôles synthétiques. Le phénotype se décrit mieux comme un continuum, ce qui exclut des sous-types discrets sans écarter une stratification par sévérité. Le lien entre données subjectives et objectives est faible, ce qui est similaire à la littérature : les coefficients les plus forts avoisinent $|\rho| = 0{,}19$ et les tailles d'effet de groupe restent inférieures ou égales à $\varepsilon^2 = 0{,}03$. Une petite composante spécifique à l'insomnie subsiste après ajustement sur la dépression et l'anxiété, se logeant dans les marqueurs circadiens et de fragmentation du sommeil. Enfin, la sévérité mesurée par l'ISI est persistante : d'une évaluation bimensuelle à la suivante, un participant conserve le même palier de sévérité avec une probabilité moyenne de 0,64, l'essentiel des changements de sévérité se limitant aux paliers adjacents.

Synthèse, limites et perspectives

Lus ensemble, ces résultats convergent. Le signal objectif qui suit la sévérité subjective de l'insomnie est faible et se situe dans les marqueurs circadiens et de fragmentation du sommeil plutôt que dans la durée de sommeil, alors que la littérature a longtemps centré cette divergence sur la durée. Les caractéristiques objectives ne se divisent pas en types discrets mais forment un continuum, et leur structure dominante est presque orthogonale aux labels d'insomnie. Le potentiel des dispositifs connectés réside dans une collecte de données non obtrusive, à domicile et longitudinale, qui donne accès à une dimension qu'une vue transversale ne peut atteindre : la manière dont un participant change d'une nuit à l'autre et au fil des semaines de suivi. Cette variabilité intra-individuelle, mesurée directement sur 26 semaines, se distingue de la moyenne du sommeil et pourrait constituer un marqueur à part entière.

Les performances du Sleep Analyzer ont été évaluées chez des adultes sains, donc son *accuracy* dans l'insomnie clinique, et avec elle la force de la mesure objective sur laquelle reposent les analyses, reste à établir. La cohorte de SLEEPI a été recrutée parmi des propriétaires de dispositifs Withings, sélectionnant ainsi une population particulière, avec certains biais, notamment de groupe socio-économique et de couleur de peau. Le label diagnostique d'insomnie applique les critères du DSM-5, fondés sur la présence de symptômes d'insomnie associés à un retentissement diurne, et en reclassant les cas comorbides en insomnie secondaire. Cependant, ce label repose sur des réponses à un auto-questionnaire plutôt que sur un entretien clinique et le critère d'opportunité adéquate de sommeil, qui n'a pas d'item dans le questionnaire, n'est donc pas appliqué. Dans l'étude réalisée dans ce projet, la prévalence de l'insomnie chronique primaire de la population étudiée est proche de 29 % dépassant ainsi largement les 6 % de la population générale répondant aux critères cliniques stricts. Enfin, le phénotypage objectif repose sur des médianes par sujet issues d'une seule coupe transversale et d'une unique version logicielle (*firmware*) d'un seul dispositif.

Les perspectives proposées découlent de ces limites. Comme l'étude n'a pas révélé de sous-groupes séparables, y compris sous une représentation non linéaire par autoencodeur, les extensions les plus prometteuses ajoutent de la supervision ou changent la nature de l'entrée : une caractérisation supervisée de l'axe de sévérité reliant les marqueurs des dispositifs aux échelles subjectives validées, et un passage des agrégats par sujet aux séquences temporelles brutes. Une piste prometteuse est l'étude de la dynamique intra-individuelle, de l'évolution de la sévérité au cours du suivi et de la comparaison de chaque participant à sa

propre référence plutôt qu'entre participants. Les algorithmes de détection veille/sommeil devraient également être évalués chez des personnes souffrant d'insomnie. La conception et la validation du protocole de l'étude clinique décentralisée à grande échelle encouragent le suivi longitudinal de différentes pathologies du sommeil, qui pourrait s'étendre par exemple à d'autres problèmes de santé, comme les pathologies cardiovasculaires ou neurodégénératives.

Cette thèse combine l'utilisation de dispositifs connectés pour le sommeil et la collecte multimodale de données sur 26 semaines à domicile, afin de caractériser une large cohorte de personnes rapportant des symptômes d'insomnie. Parce que le profil objectif de l'insomnie forme un continuum plutôt que des classes nettement délimitées, un patient pourrait être mieux situé le long d'un gradient de sévérité qu'assigné à un phénotype discret. À plus long terme, et si les marqueurs objectifs et leurs trajectoires se confirment, l'ISI combiné aux variables issues des dispositifs pourrait aider à décider d'une orientation et à adapter la prise en charge au patient. Le phénotypage de l'insomnie à partir de données de capteurs en conditions réelles demeure le principal problème ouvert de ce travail.

Acknowledgments

This adventure made me realize how fortunate I was to be surrounded by people who wished me well and believed I would succeed.

My first thanks go to my supervisors. To Guillaume, for your trust, your support, and your everyday kindness: you are an office neighbor and foosball teammate whose company is a genuine pleasure. Kæra Anna, hjartans þakkir fyrir góðvild þína, auðmýkt og óbilandi hughreystingu, og fyrir að hafa kynnt mig fyrir þessu yndislega landi. To Pierre, for your trust and your availability.

I also thank Martin de Gourcuff, without whom the idea of doing a PhD would never have crossed my mind, and who allowed me and pushed me to be ambitious for this project; and David Campo, for his time and his rigor, which did nothing to soften my perfectionism. I am grateful to Estelle Durand for her important role on my thesis monitoring committee, where she established a climate of trust that allowed open communication.

To my colleagues at Withings, whom I now count as friends: to Vincent, for his rigor and making sure I never ran out of chocolate in hard times; to Pampam, an unbeatable tram and concerts companion and always willing to help me; to Violaine, for arriving in the right place at the right time; but also to Léa, Constance, Jiachuan, Louis, Mahot, Laura, more broadly, to the whole Machine Learning team, and to all those I do not name, who followed the highs and lows of my adventure with kindness, support, and empathy.

To the laboratories that welcomed me: to the Sleep Revolution team, for making me feel at home when I was closer to Greenland than to my own, and to ChronoS, for the good cheer and the great company during breaks.

To my friends: to Benjamin, for our regular meet-ups, when I had given up hope that our paths would cross geographically again; to Muneeb, because you believed in me when I thought I could not make it; to Noelia, Arthur, Jonathan, the two Cocos, Nathan and Léa, and to all the friends I do not name: the omission says nothing about my gratitude.

À ma famille, pour votre patience pendant ces longues études qui arrivent à leur terme.

À Boti, de tout mon cœur : merci d'avoir pleinement vécu cette aventure riche et intense à mes côtés, même quand tout tournait autour de nos thèses respectives. Merci d'avoir été mon plus grand soutien dans les moments difficiles, professionnels comme personnels. Merci de m'avoir redonné confiance en moi quand j'en manquais. Et merci d'avoir toujours décrit ma thèse avec justesse et fierté à tes proches.

Finally, I thank Withings for supporting this project financially and, above all, for allowing me to work and exchange with so many passionate people who agreed to lend a hand and share their expertise, and without whom this ambitious project could not have come to be.

And, lastly, to the project, for making me grow as a researcher, for the many lessons it has taught me about science and about life, and for giving me the opportunity to really understand what it is like to suffer from insomnia, but also for teaching me how to have good nights' sleep!

Contents

Abstract i

Résumé iii

Résumé substantiel en français v

Acknowledgments xi

List of figures xx

List of tables xxii

List of abbreviations xxiii

1 Introduction 1
- 1.1 Background and rationale . . . 1
- 1.2 Insomnia . . . 2
  - 1.2.1 Definitions . . . 2
  - 1.2.2 Insomnia disorder versus insomnia complaint . . . 2
  - 1.2.3 Pathophysiology and causes . . . 3
  - 1.2.4 Diagnosis and treatment . . . 3
- 1.3 Digital tools for sleep assessment . . . 4
  - 1.3.1 Wearables and nearables . . . 4
  - 1.3.2 Machine learning for insomnia detection . . . 5
- 1.4 Goals . . . 7

2 Digital solutions for insomnia assessment: systematic review 9
- 2.1 Rationale . . . 9
- 2.2 Objectives . . . 10
- 2.3 Methods . . . 10
  - 2.3.1 Search strategy . . . 10
  - 2.3.2 Deduplication . . . 11
  - 2.3.3 Screening process description . . . 14
  - 2.3.4 Initial scope definition and scope refinement: insomnia and sleep quality . . . 15
  - 2.3.5 Data extraction process . . . 15
- 2.4 Results . . . 15

2.4.1 Screening results . . . 15
2.4.2 Datasets . . . 16
2.4.2.1 Publicly available datasets vs. proprietary datasets . . . 16
2.4.2.2 Assessment of differences across datasets . . . 17
2.4.3 Methods . . . 26
2.4.3.1 Input data . . . 26
2.4.3.2 Models and outputs . . . 33
2.4.3.3 Validation method . . . 35
2.5 Limitations . . . 36
2.5.1 Search strategy . . . 36
2.5.2 Screening and selection . . . 36
2.5.3 Data extraction . . . 37
2.5.4 Scope and coverage . . . 37
2.6 Conclusion . . . 37
2.6.1 Synthesis of the evidence landscape . . . 37
2.6.2 Feature design . . . 38
2.6.3 Independence of the evidence base . . . 38
2.6.4 Directions for future research . . . 38
2.6.5 Toward phenotyping with multi-night home monitoring . . . 39
2.7 Chapter Summary . . . 40

3 Evaluation of the Sleep Analyzer's sleep stage identification algorithm 41
3.1 Introduction . . . 41
3.2 Paper . . . 41
3.2.1 Evaluation of a Contactless Sleep Monitoring Device for Sleep Stage Detection at Home in a Healthy Population: Prospective Study in Free-Living Conditions . . . 42
3.2.1.1 Abstract . . . 42
3.2.1.2 Introduction . . . 43
3.2.1.3 Methods . . . 44
3.2.1.4 Results . . . 46
3.2.1.5 Discussion . . . 55
3.2.1.6 Conclusions . . . 59
3.2.1.7 References . . . 60
3.3 Chapter Summary . . . 65

4 Data Collection: SLEEPI Protocol 67
4.1 Introduction . . . 67
4.2 Paper . . . 67
4.2.1 End-to-end Protocol for large-scale, digital, longitudinal and decentralized clinical trials: the SLEEPI study on insomnia patterns . . . 68
4.2.1.1 Abstract . . . 68
4.2.1.2 Introduction . . . 70
4.2.1.3 Methods . . . 71
4.2.1.4 Enrollment Status . . . 84
4.2.1.5 Discussion . . . 85
4.2.1.6 Conclusion . . . 86

4.2.1.7 References . . . 88
4.3 Chapter Summary . . . 95

5 Subjective and objective characterization of insomnia in the SLEEPI cohort 97
5.1 Data quantity . . . 97
5.2 Demographics . . . 99
5.3 Definition of the insomnia labels . . . 100
5.3.1 Insomnia diagnosis classes . . . 100
5.3.2 Insomnia severity . . . 102
5.4 Subjective clinical data . . . 103
5.4.1 Descriptive statistics . . . 103
5.4.2 Correlations . . . 103
5.4.3 Comorbidity . . . 106
5.5 Objective data . . . 106
5.5.1 Data segmentation . . . 106
5.5.2 Feature building . . . 107
5.6 Clustering . . . 108
5.6.1 Model Selection . . . 108
5.6.1.1 Traditional partitioning . . . 110
5.6.1.2 Non-linear autoencoder . . . 110
5.6.2 Results . . . 110
5.6.2.1 Traditional partitioning . . . 110
5.6.2.2 Non-linear autoencoder . . . 111
5.6.2.3 Interpretation . . . 113
5.6.2.4 Night-to-night variability . . . 113
5.6.2.5 Beyond a cross-sectional partition . . . 114
5.7 Subjective–objective relationship . . . 114
5.8 Chapter Summary . . . 118

6 Discussion 121
6.1 Reading the findings together . . . 121
6.2 Limitations . . . 124
6.3 Perspectives . . . 124
6.4 Conclusion . . . 125

A Supplementary material for Chapter 2 (systematic review) 127
A.1 Detailed search queries . . . 127
A.2 Records included in the full analysis . . . 129
A.3 Detailed inclusion and exclusion criteria . . . 135
A.3.1 Common exclusion domains . . . 135
A.3.2 Insomnia-group-specific criteria . . . 136
A.4 Detailed subjective-data instruments . . . 137
A.4.1 Public-dataset studies . . . 137
A.4.2 Proprietary-dataset studies . . . 137
A.5 Detailed recruitment channels . . . 138
A.6 Detailed alternative recording tools . . . 139
A.7 Detailed objective-data modalities and sampling fidelity . . . 139

A.8 Detailed monitoring duration per subject . . . 140
A.9 Night and diurnal data by study . . . 141
A.9.1 Publicly available datasets . . . 141
A.9.2 Proprietary datasets . . . 141
A.9.3 Classification notes . . . 141
A.10 Detailed validation designs . . . 142
A.10.1 Supervised model validation . . . 142
A.10.1.1 Resampling . . . 142
A.10.1.2 Nested resampling with held-out test set . . . 142
A.10.1.3 Single train/test partition . . . 143
A.10.1.4 Cross-dataset validation . . . 143
A.10.2 Unsupervised cluster validation . . . 143
A.10.2.1 External cluster validity against independent labels . . . 144
A.10.2.2 No validation reported . . . 144
A.10.3 Statistical inference and signal-processing studies . . . 144
A.11 Detailed models and outputs . . . 145
A.11.1 Binary classification . . . 145
A.11.2 Multi-class and multi-task classification . . . 146
A.11.3 Cluster-based phenotypic stratification . . . 147
A.11.4 Statistical and signal-level outputs . . . 148
A.12 Detailed feature extraction . . . 148
A.12.1 Variants of traditional parameters . . . 148
A.12.2 Mathematical transformations and signal decomposition . . . 149
A.12.3 Neuroimaging features . . . 150
A.12.4 Subjective instruments . . . 150
A.13 Detailed preprocessing steps . . . 152

B Supplementary material for Chapter 3 (at-home Sleep Analyzer evaluation study) 155
B.1 Details on methods and algorithm performance per sleep stage . . . 155
B.2 Results of Kruskal-Wallis and Brown-Forsythe statistical tests followed by Holm-Bonferroni correction . . . 179
B.3 Comparison with previous research . . . 182

C Supplementary material for Chapter 4 (SLEEPI data-collection protocol) 185
C.1 Identification and assessment of data loss sources . . . 185
C.2 Prospective Attrition Model . . . 187
C.3 A priori power analysis for sample-size determination . . . 189

D Supplementary material for Chapter 5 (SLEEPI data analyses) 195
D.1 Device-screening exits . . . 195
D.2 Subjective clinical data: supplementary figures . . . 195
D.3 Objective feature definitions . . . 196
D.4 Objective data: supplementary figures . . . 198
D.5 Clustering: supplementary figures . . . 198
D.6 Subjective–objective relationship: supplementary figures . . . 199

References **207**

List of Figures

1.1 The two Withings devices studied in this thesis 5

2.1 Search strategy combining three semantic domains. 11
2.2 PRISMA flow diagram of the systematic review process 13
2.3 Monitoring duration per subject across datasets 19
2.4 Night and diurnal data capture across datasets 20
2.5 Included studies by primary output type 34

Ch. 3 paper, Figure 1: Study inclusions flowchart 45
Ch. 3 paper, Figure 2: Example hypnogram, WSA sleep stages versus the PSG consensus 45
Ch. 3 paper, Figure 3: Sleep-stage classification confusion matrix 48
Ch. 3 paper, Figure 4: RANSAC regression of WSA versus PSG total sleep time . 49
Ch. 3 paper, Figure 5: Bland–Altman plot, total sleep time and time in REM . . . 50
Ch. 3 paper, Figure 6: Bland–Altman plot, time in deep and light sleep 51
Ch. 3 paper, Figure 7: TST mean absolute error by age group 51
Ch. 3 paper, Figure 8: TST mean absolute error by BMI group 52
Ch. 3 paper, Figure 9: TST mean absolute error by bed-partner group 52
Ch. 3 paper, Figure 10: Confusion matrix between the two reviewers' annotations 52
Ch. 3 paper, Figure 11: Confusion matrices on agreement and disagreement sets 53
Ch. 3 paper, Figure 12: Reported sleep quality with versus without PSG 54
Ch. 3 paper, Figure 13: Sleep-wake sensitivities across validation studies 56
Ch. 3 paper, Figure 14: Sleep-stage sensitivities across validation studies 57

Ch. 4 paper, Figure 1: Timing and sequence of study questionnaire administrations 72
Ch. 4 paper, Figure 2: Schedule of nudges for incomplete participant tasks 75
Ch. 4 paper, Figure 3: Comparison of actual and projected participant inclusions 79
Ch. 4 paper, Figure 4: Distribution of participant age 79
Ch. 4 paper, Figure 5: Distribution of participant-owned devices 80
Ch. 4 paper, Figure 6: Illustration of participant drop-off points 80
Ch. 4 paper, Figure 7: Power curves for insomnia classification under different AUC criteria 81
Ch. 4 paper, Figure 8: Illustration of the prospective attrition model 82

5.1 Initial-survey completion by section 98
5.2 Recurring-ISI participation by iteration 98
5.3 Distributions of height, weight, and age across the analyzable sample. 100

5.4 Self-reported skin color . . . 101
5.5 Self-reported socio-economic category across the analyzable sample. . . . 101
5.6 Self-reported ethnicity, a multiple-choice question, across the analyzable sample. . . . 102
5.7 DSM-5 insomnia prevalence by survey iteration . . . 103
5.8 ISI by insomnia subtype . . . 105
5.9 Distributions of sleep, cardiac, activity and circadian features . . . 109
5.10 PCA of the objective feature set . . . 109
5.11 PCA and UMAP projections by DSM-5 class . . . 111
5.12 UMAP colored by feature values and mean ISI . . . 112
5.13 PCA and UMAP of the autoencoder latent by ISI tier . . . 112
5.14 Per-participant ISI trajectories . . . 115
5.15 Wave-to-wave ISI tier transitions . . . 116
5.16 Partial Spearman correlation of features with ISI . . . 117

6.1 Overview of the thesis contributions . . . 122

D.1 Depression and anxiety prevalence by ISI tier . . . 195
D.2 Co-occurrence of insomnia, depression and anxiety . . . 199
D.3 Device coverage by ISI tier and DSM-5 class . . . 200
D.4 Spearman correlation among objective features . . . 200
D.5 Missingness in the objective feature matrix . . . 201
D.6 Distributions of the circadian metrics . . . 201
D.7 Internal cluster-validity scores . . . 202
D.8 Feature profile of the best two-group partition . . . 202
D.9 Feature–score Spearman correlations, full screen . . . 203
D.10 Objective sleep features by DSM-5 class . . . 204
D.11 Objective sleep features by ISI tier . . . 204
D.12 Correlations of variability features with scores . . . 205
D.13 Sleep duration and efficiency by complaint dimension . . . 206
D.14 mRMR feature ranking per score . . . 206

List of Tables

1.1 The four research axes of the thesis 7

2.1 Summary of databases, search queries, and reference counts 12
2.2 Criteria for inclusion and exclusion in the systematic review 14
2.3 Overview of publicly available datasets used across selected studies 16
2.4 Demographic and clinical characteristics of included datasets 17
2.5 Objective signal modalities across dataset records 18
2.6 Representative exclusion criteria across studies 21
2.7 Subjective instruments across studies 23
2.8 Recording tools beyond actigraphy and PSG 25
2.9 Traditional sleep and activity parameters used across insomnia datasets . . 28
2.10 Mathematical data transformations and their clinical intuition 30
2.11 Included studies by validation paradigm 35

Ch. 3 paper, Table 1: Demographic and clinical characteristics of the population 46
Ch. 3 paper, Table 2: Summary of sleep-stage classification results over nights . 48
Ch. 3 paper, Table 3: Bland–Altman results on sleep-related metrics 50
Ch. 3 paper, Table 4: Sleep-stage episodes and stage changes, WSA versus PSG . 50
Ch. 3 paper, Table 5: Performance in reviewers' agreement and disagreement subgroups 53
Ch. 3 paper, Table 6: Reported sleep quality and sensor disturbance with versus without PSG 54

Ch. 4 paper, Table 1: Summary of study validation and phased implementation . 84

5.1 Per-participant data-quantity metrics 99
5.2 Continuous demographic variables 100
5.3 Distribution of the subjective clinical scores 103
5.4 Correlations among the subjective scores 104
5.5 ISI score by DSM-5 insomnia class 105
5.6 Depression and anxiety prevalence by DSM-5 class 106
5.7 Distribution of the main objective features 108
5.8 Key results of the SLEEPI analyses 118

A.1 Database platforms, formats, and search queries 128
A.2 The 49 records included in the full analysis of the systematic review. 130

D.1 Device-screening exits by failure mode 196

D.2 Definitions of the 39 objective features. 196

List of abbreviations

AASM American Academy of Sleep Medicine. 33, 34, 130

ACM Association for Computing Machinery. 16, 19, 112

AHI Apnea-Hypopnea Index. 36

BMI Body Mass Index. 28, 34, 114

CAE Convolutional Autoencoder. 40, 41

CAP Cyclic Alternating Pattern. 11, 12, 46, 118

CBT-I Cognitive Behavioral Therapy for Insomnia. 8

COMISA Comorbid Insomnia and Sleep Apnea. 118

CV Cross-Validation. 44, 120–122

DSM-5 Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition. 7, 8, 13, 32, 47, 119

ECG electrocardiogram. 9, 11–13, 24, 31, 35, 117, 121

EEG electroencephalogram. 9, 11–13, 24, 33, 35, 39, 41, 45, 46, 48, 109, 121, 122, 124, 125, 128, 130, 131

EMA Ecological Momentary Assessment. 47

EMG electromyogram. 9, 24, 121

EOG electrooculogram. 9, 24, 121

HR Heart Rate. 9

HRV Heart Rate Variability. 9, 122

ICSD-3 International Classification of Sleep Disorders, Third Edition. 8

Inf-FS Infinite Feature Selection. 39–41

IS Interdaily Stability. 35, 36

ISI Insomnia Severity Index. 8, 12, 13, 15, 27, 129

IV Intradaily Variability. 35, 36

LASSO Least Absolute Shrinkage and Selection Operator. 40

LCA Latent Class Analysis. 39, 41

LOO Leave-One-Out. 44, 121

LOSO Leave-One-Subject-Out. 44, 121, 122

MASS Montreal Archive of Sleep Studies. 122

MEDLINE Medical Literature Analysis and Retrieval System Online. 19

MRI magnetic resonance imaging. 11, 12

mRMR Minimum Redundancy Maximum Relevance. 39, 41

PCA Principal Component Analysis. 39–41, 128

PPG photoplethysmography. 10

PRISMA Preferred Reporting Items for Systematic reviews and Meta-Analyses. 16, 21, 45

PSG polysomnography. 3, 8, 9, 11–13, 34–37, 39–41, 127, 128

PSQI Pittsburgh Sleep Quality Index. 15, 27, 129

REM rapid eye movement. 124

RFE Recursive Feature Elimination. 40

RR respiratory rate. 9

SD standard deviation. 37

SE sleep efficiency. 15, 21

SL sleep latency. 21

SpO2 oxygen saturation. 24

SVM Support Vector Machine. 48, 121

SW ScanWatch. 10, 109

SW 2 ScanWatch 2. 10, 118

t-SNE t-distributed Stochastic Neighbor Embedding. 40

TST total sleep time. 8, 12, 15, 21

WASO wake time after sleep onset. 8, 12, 15, 21, 29, 115, 130

WSA Withings Sleep Analyzer. 10, 13, 14, 109

Chapter 1

Introduction

The quality of sleep varies from one night to the next [1] and differs between individuals. The reference sleep laboratory examination, polysomnography (PSG), records a single night and therefore does not follow these variations over time. Wearable and nearable devices now measure sleep at home over long periods, which makes it possible to record how sleep changes from night to night [2]. This continuous measurement creates new opportunities to characterize insomnia, which is the focus of this thesis [3, 4].

1.1 Background and rationale

Insomnia affects 16% of the population worldwide [5] and 19% of the French population [6], making it the most common sleep disorder according to the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) [7]. A 2025 French national survey reported that 43% of adults experience at least one sleep disorder [8], with women being more affected than men [9]. These proportions, just as other figures linked to insomnia detection, vary depending on the diagnostic criteria used. This pathology affects the quality of life of individuals with insomnia on multiple dimensions [10]: physical, social functioning, emotion, and mental health [11, 12]. Insomnia is associated with an increased risk of accidents due to reduced alertness and cognitive functions, and with the later development of metabolic and cardiovascular disease [12, 13]. Whether insomnia is consistently associated with an increased risk of adverse health outcomes such as cardiometabolic disease is still debated, as prospective studies disagree, some reporting a higher risk of heart failure or stroke and others none [14]. Two factors explain much of this disagreement. The risk depends on how insomnia is defined and on the symptom involved: trouble falling asleep raised the risk of heart disease and stroke by 27% and unrefreshing sleep by 18%, while waking too early carried no clear risk [14]. Many studies also adjust for factors such as short sleep and depression that may sit on the path from insomnia to disease, which can hide a real effect; studies using Mendelian randomization avoid this and still link insomnia to coronary artery disease, heart failure, and stroke [15]. The risk is concentrated in people whose sleep is objectively short: this phenotype shows up to a fivefold higher odds of

hypertension, about threefold for type 2 diabetes, and higher mortality [16], and objective short sleep roughly doubled the risk of new hypertension whereas feeling one sleeps little did not [17]. Insomnia with normal objective sleep carries no such risk and more often resolves. Identifying these at-risk phenotypes is a central aim of this thesis, so that the people most likely to be harmed, or to benefit from treatment, can be found. The burden increases when insomnia co-occurs with major depression, a combination associated with poorer functioning, lower quality of life, and higher healthcare costs across five European countries [18].

1.2 Insomnia

1.2.1 Definitions

The main definitions of insomnia (DSM-5 [7] and International Classification of Sleep Disorders, Third Edition (ICSD-3) [19]) are based on difficulties falling asleep, sleep maintenance, too early waking, the frequency and duration of these disorders, and the diurnal consequences of these sleep troubles.

1.2.2 Insomnia disorder versus insomnia complaint

Insomnia is a heterogeneous pathology defined by the patient's self-reported complaint of poor sleep. Objective metrics from actigraphy or baseline polysomnography (PSG), such as SL and wake time after sleep onset (WASO) [20], provide adjunct measurements rather than defining the disorder, and studies show that the difference between people with insomnia and regular sleepers is primarily characterized by differences in total sleep time (TST) [21]. The severity of symptoms is measured using questionnaires such as the Insomnia Severity Index (ISI) [22, 23, 24, 25], whose French version has been validated [26], especially for patients already diagnosed with treatment such as Cognitive Behavioral Therapy for Insomnia (CBT-I) [27, 28]. Physiological recordings and the patient's self-reported perception of poor sleep do not always coincide. Indeed, the two correlate poorly: while 13% of good subjective sleepers meet objective short-sleep criteria [21], a large proportion of individuals in the general population report being bothered by sleep problems, reaching 65% in France, 77% in the USA, 53% in Japan, and 50% in Italy [29], while only about 6% meet the strict diagnostic criteria for chronic insomnia [30]. These figures illustrate the scale of the problem, from a widespread complaint to a much smaller clinical disorder. In order to improve the care trajectory and more personalized management of individuals with insomnia symptoms, it appears necessary to use new longitudinal means of evaluation integrating both subjective and objective markers.

The sleep literature conventionally refers to study populations as *patients*. The populations considered in this thesis, and in the datasets surveyed in the systematic review of Chapter 2, are broader than this term suggests: they comprise clinically diagnosed patients, participants enrolled in clinical trials who present insomnia symptoms without necessarily

having been diagnosed, and people with no sleep complaint included as controls. The terms *people with insomnia symptoms* and *individuals with insomnia symptoms* are therefore used interchangeably in this work, instead of *patients*.

1.2.3 Pathophysiology and causes

Insomnia has no single cause. Its most consistent feature is hyperarousal, a persistent overactivity of the brain's arousal systems that is present during both day and night and spans physiological, cognitive, and emotional processes, with raised autonomic activity and higher output of the stress hormone cortisol through the hypothalamic-pituitary-adrenal axis [31, 32]. How the disorder becomes chronic is described by the 3P model: predisposing traits and a precipitating stressor set it off, and perpetuating factors then keep it going after the original trigger has faded, including arousal that becomes conditioned to the bed, worry and self-monitoring of sleep, and spending excessive time in bed [33, 31]. The predisposition is substantially genetic and highly polygenic, with many genes of small effect and heritability estimates between about 22% and 59% in adults; the risk genes concentrate in circuits involved in emotion regulation and are shared with psychiatric traits, so genetics act as a final common pathway rather than a single cause [34, 31]. Circadian and sleep-homeostatic processes contribute but are not the primary cause in most patients [31]. Instability of rapid eye movement (REM) sleep has also been proposed to disturb the overnight processing of emotions and to sustain both the misperception of sleep and the persistence of hyperarousal [31].

1.2.4 Diagnosis and treatment

Insomnia is under-recognized and under-treated: in France, although 73% of individuals with insomnia report sleep-maintenance difficulties and 61% difficulty falling asleep, only 22% have consulted a physician about it and 14% have received treatment [29]. A French Delphi consensus provides non-sleep-specialist physicians with a structured approach to managing insomnia complaints in primary care [35].

Among objective tools, actigraphy tracks sleep-wake and circadian patterns at low cost over long periods, but the European guideline finds only limited support for its use in diagnosing insomnia, since it cannot reliably distinguish individuals with insomnia from good sleepers [36]; it estimates sleep from movement, does not resolve sleep architecture, and overestimates sleep by scoring quiet wakefulness as sleep, a phenomenon that is pronounced in individuals with insomnia who lie awake without moving [37, 38]. PSG is the reference examination for objectively identifying sleep abnormalities through the signals it measures: electroencephalogram (EEG), electrooculogram (EOG), electromyogram (EMG), electrocardiogram (ECG), pulse oximetry, air flow and breathing effort [39]. If this medical examination is performed, the PSG takes time to set up and has a high cost for routine use, reaching 930€ for a sleep day-hospital PSG session at Bichat APHP (Paris, France) in 2022. Moreover, the scoring of those signals is manual labor with risk of human errors and long queues, since there are only few specialized scorers per country. In addition, PSG is also

only applied for one or two nights so it cannot reflect circadian rhythms and sleep/wake cycle variability nor can it account for variability between nights [1]. The laboratory environment can also distort the measurement in the opposite direction: many people with insomnia sleep better in the sleep laboratory than at home, a reversal of the usual first-night effect, because in conditioned insomnia the home bedroom has become a learned cue for arousal that the unfamiliar laboratory does not carry [40]. A single laboratory recording can therefore underestimate the sleep disturbance that occurs at home, which strengthens the case for monitoring sleep at home across several nights. Reliable estimates of habitual sleep in fact require many nights even with a consumer tracker: roughly three to five nights for a good weekly total-sleep-time estimate, about ten for a monthly one, and eleven to eighteen to capture night-to-night variability [41]. PSG is not routinely indicated for insomnia, being reserved for suspected comorbid sleep disorders, treatment-refractory cases, and a marked subjective–objective discrepancy [36]. Sleep apnea frequently co-occurs with insomnia: about 35% of people with insomnia also have obstructive sleep apnea, and about 38% of people with obstructive sleep apnea also have insomnia, a combination known as Comorbid Insomnia and Sleep Apnea (COMISA) [42]. COMISA is harder to treat than either disorder alone, because the insomnia lowers acceptance of and adherence to continuous positive airway pressure, so current guidance is to treat the insomnia first [43]. This frequent overlap is a reason to screen for sleep apnea when assessing insomnia.

Additional parameters such as Heart Rate (HR), Heart Rate Variability (HRV) and respiratory rate (RR) and variability should be more discriminative to help derive sleep/wake scoring [44, 45, 46]. Across traditional measurement tools and smart devices, sleep and insomnia are inferred from physiological signals, including brain activity, cardiac and respiratory activity, body movement, sound, and pressure, which constitute the inputs analyzed by detection algorithms [39, 4].

1.3 Digital tools for sleep assessment

1.3.1 Wearables and nearables

In order to improve the management of people suffering from insomnia, the introduction of smart devices requiring little or no physical contact with the user, whether wearable or nearable, is a growing need. Based on actigraphy and some vitals estimation, such as HR and RR, smart devices enable continuous long-term monitoring of sleep at home [47]. The democratization of smart devices for sleep would improve access to lower-cost care than current actigraphy and PSG examinations, and would gather signals measured in the patients' real sleep conditions, since laboratory-based PSG fails to represent their daily lives or their sleep-wake rhythms and gives no measure of intra-individual variability. Moreover, gathering longitudinal data has the potential to yield novel insights into fluctuations in insomnia symptoms over time [1]. Besides, these new tools can generate new scientific and medical knowledge to characterize the subtypes of insomnia and identify new therapeutic targets [3, 4].

Digital solutions, such as wearables and nearables paired with dedicated applications,

support multimodal approaches for early symptom detection and prevention [48] thanks to data collection with multiple biomarkers and the development of new objective tools combined with subjective polls.

Withings develops connected health devices that record physiological signals during everyday life and present the resulting analysis to the user through a mobile application. This thesis focuses on two Withings devices selected for their sleep tracking capabilities and broad adoption: Withings Sleep Analyzer (WSA) [49], a nearable device, and ScanWatch (SW) [50], a wearable device. On the one hand, the WSA (Figure 1.1a), which is placed under the mattress of the bed, contains a pressure sensor as well as a microphone, and allows the analysis of both sleep architecture and ventilatory parameters during sleep. For instance, multi-night studies have shown that the Apnea-Hypopnea Index (AHI) it measures varies substantially from night to night, so that the repeated, ecological recordings it provides are medically valid and more informative than a single laboratory night [2]. On the other hand, the SW (Figure 1.1b) is an enriched actigraphy smartwatch composed of an accelerometer and a photoplethysmography (PPG) sensor [51] whose signals enable us to deduce information such as HR and oxygen saturation [52]. The SW allows the measurement of sleep-wake rhythms and multiple biomarkers of circadian rhythms. The ScanWatch 2 (SW 2) additionally incorporates a body temperature sensor.



(a) Withings Sleep Analyzer (WSA) (b) ScanWatch 2 (SW 2)

Figure 1.1: The two Withings devices studied in this thesis: the Withings Sleep Analyzer (WSA) (left), a nearable placed under the mattress, and the ScanWatch 2 (SW 2) (right), a wearable smartwatch.

1.3.2 Machine learning for insomnia detection

Existing heuristics used to assess insomnia disorder [7] remain limited: they do not match insomnia complaints [29] and are most often not built on a temporal window long enough to follow insomnia symptoms over time and to capture inter-subject variability. Machine learning methods that incorporate many parameters are therefore promising tools for greater

generalization in insomnia assessment.

Existing studies rely on physiological and behavioral inputs. Actigraphy alone, or EEG and ECG from the PSG measurement, are the most common signals [53, 54, 55, 56], as surveyed in Chapter 2. Some work uses magnetic resonance imaging (MRI) to investigate insomnia mechanisms together with questionnaires, while others add sleep patterns and daytime activities collected through continuously worn smart watches [57].

Four algorithmic families have been explored. Most studies perform supervised classification on hand-crafted features, whether derived from functional connectivity [58] or from actigraphy-derived statistical descriptors and clinical sleep parameters such as TST and WASO [59], with the SVM as the most frequent classifier and the random forest or neural network as the second most common. In contrast, clustering-based approaches operate directly on raw time series rather than extracted features [60]. One approach deduces a subject's insomnia status after segmenting the night into sleep stages [61], as done in one study [62]. Other studies instead classify segments of the night individually and infer the subject's insomnia status from the proportion of segments classified as insomnia disorder [53, 54, 55]. No study, however, evaluates the night as a whole, which could provide additional information for insomnia assessment.

A recurring limitation across these studies is the amount of available data, which makes it difficult both to train models that generalize reliably and to assess their generalization on held-out data. A subset of the publicly available Cyclic Alternating Pattern (CAP) dataset [63] contains only 13 subjects, equally split between regular sleepers and individuals with insomnia, each recorded for a single night during a PSG. The unfamiliar laboratory setting and the recording sensors degrade sleep on a first recording night, a phenomenon known as the first-night effect [64], so this single night can make a regular sleeper appear to sleep poorly and thereby blur the distinction between regular sleepers and individuals with insomnia symptoms. A larger internal dataset of 80 subjects equally balanced between controls and individuals with insomnia was used for a deep learning algorithm applied to 2-channel EEG signals [53], but remains too small to yield a reliable estimate of generalization performance.

Beyond dataset size, the design of evaluation protocols also warrants attention. For instance, the study [55] of binary classification for insomnia detection from 30-second night ECG signals reports a 97% accuracy; however, the training and evaluation splits are defined at the segment level rather than at the subject level, so segments from the same individual may appear in both sets, which tends to inflate performance estimates. A hold-out CV at the subject level performed in a second stage yields a 94% accuracy but on a dataset too small to guarantee representativeness. In contrast, the deep learning study on 2-channel EEG [53] evaluates its model with 5-fold CV and tunes it on a separate fixed validation set of only 8 subjects, avoiding data leakage, and a later study by the same author [54] adopts LOO CV. These designs remain constrained by dataset size. Overall, datasets large enough to cleanly separate training, validation and testing sets are required for reliable machine learning methods [65].

Ground-truth insomnia labels are another weakness. Public datasets [63, 66] and proprietary datasets [67, 46, 68] rarely document how insomnia was assessed; the ISI [22] is a plausible choice but is seldom explicitly reported. Labels could be strengthened by combin-

ing DSM-5 criteria [7] with additional questionnaires [22, 23, 24, 25]. Moreover, a one-night PSG is itself prone to sleep disturbance and may bias the reference diagnosis.

1.4 Goals

This thesis aims to characterize insomnia and to search for homogeneous subgroups of the disorder through four research questions, each answered in a dedicated chapter and prepared as an independent publication. Table 1.1 summarizes these four axes together with the chapter that addresses each and its publication status.

Table 1.1: The four research axes of the thesis, with the chapter answering each research question and the associated publication status.

#	**Research question**	**Contribution**	**Publication status**
1	What is the state of the art of digital solutions for insomnia assessment?	Systematic review of digital tools for insomnia detection and characterization (Chapter 2).	To be submitted, *Sleep Medicine Reviews*
2	How accurately can a nearable identify sleep stages in real-world conditions?	Evaluation of the Withings Sleep Analyzer sleep-stage algorithm against polysomnography in an at-home study (Chapter 3).	Published, *JMIR Human Factors*
3	Can a longitudinal multimodal dataset combining objective sensor recordings and subjective questionnaire responses be collected ecologically?	Design and execution of the SLEEPI data-collection protocol (Chapter 4).	Under review, *JMIR* group
4	Can homogeneous subgroups of individuals with insomnia be identified from longitudinal sleep-pattern evolution?	Prospective analysis of longitudinal Withings device data, testing for objective subgroups and quantifying within-person variability (Chapter 5).	To be submitted

Chapter 2

Digital solutions for insomnia assessment: systematic review

2.1 Rationale

Insomnia is diagnosed solely from the subjective complaint, as defined by the DSM-5 [7] and the ICSD-3 [19], with severity assessed through questionnaires such as the ISI and the Pittsburgh Sleep Quality Index (PSQI) [36]. Actigraphy and PSG provide objective measurements that may corroborate or diverge from this complaint without defining the disorder. Objective metrics such as TST, WASO, and sleep efficiency (SE) frequently fail to correlate with the subjective experience [21], a divergence that likely stems from multiple factors, including sleep state misperception, the ecological limitations of clinical assessment, and the recording of objective parameters that may not capture the features underlying the subjective complaint. Objective measures therefore do not currently contribute to the clinical diagnosis, which rests on the subjective complaint, but they may add value if they can identify insomnia phenotypes that are amenable to targeted treatment or that carry a higher risk of harm [16]. This prospect motivates examining tools that combine objective and subjective information rather than either alone.

Among objective methods, actigraphy tracks sleep-wake and circadian patterns and their night-to-night variability at low cost over long periods, but the European guideline finds only limited support for its use in diagnosing insomnia, since it cannot reliably distinguish individuals with insomnia from good sleepers [36]; it estimates sleep from movement, does not resolve sleep architecture, and overestimates sleep by scoring quiet wakefulness as sleep, which is pronounced in individuals with insomnia who lie awake without moving [37, 38]. PSG provides the highest-resolution measurement but is obtrusive, expensive, and requires configuration by trained professionals; a single night cannot reproduce natural sleep conditions or capture night-to-night variability [1], and it is not routinely indicated for insomnia, being reserved for suspected comorbid sleep disorders, treatment-refractory cases, and a marked discrepancy between subjective and objective sleep [36].

Digital tools including smart devices and mobile applications offer the potential to sup-

port long-term analysis for improved insomnia symptom detection, early intervention, and management. Consequently, a systematic review of existing digital tools such as contactless devices, wearables, applications, and standalone algorithms capable of detecting or characterizing insomnia will guide hardware development, clinical validation, and the integration of connected devices into clinical practice and public health initiatives. While systematic reviews on similar topics exist, none focuses specifically on digital tools for insomnia. Some are too restrictive, addressing digital cognitive behavioral therapy [69], machine learning methods [57], or sleep-wake disorders [70]; others are too broad, covering general sleep monitoring [4], photoplethysmography [71], or automated detection across all sleep disorders [72]. The closest prior work reviews automated insomnia detection from physiological signals [73, 74] or objective sleep assessment for insomnia management [75], but these concentrate on signal-based algorithms or on clinical management, and none examines the objective-and-subjective combination this review targets.

2.2 Objectives

Which digital solutions are available for insomnia detection and characterization, including wearable and nearable devices, their embedded and standalone algorithms, and mobile applications?

This literature review aims to understand which data insomnia ground truth and relevant features are defined on, and what the existing insomnia screening methods are. To do so, we compared the existing modalities and methods used for identifying objective biomarkers of insomnia disorder. The scope includes all potential indicators relevant to insomnia screening and phenotyping.

2.3 Methods

This section details the systematic review methodology used to identify and evaluate digital solutions for insomnia assessment. This literature review was guided by the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines [76]. The review protocol was registered in PROSPERO prior to the search under identifier CRD42024512504. The review process encompassed database selection, search strategy formulation, deduplication, screening, and data extraction. The overall process is illustrated in Figure 2.2.

2.3.1 Search strategy

As the goal was to identify digital solutions for insomnia assessment from signals of any nature, the query combined three components, namely digital solutions, insomnia, and input signals, joined by the boolean operator AND, with each component expanded into keywords and synonyms joined by OR, as shown in Figure 2.1. Major bibliographic databases

were queried through four online platforms: PubMed, Web of Science, Scopus, and the Association for Computing Machinery (ACM) Digital Library, as detailed in Table 2.1.

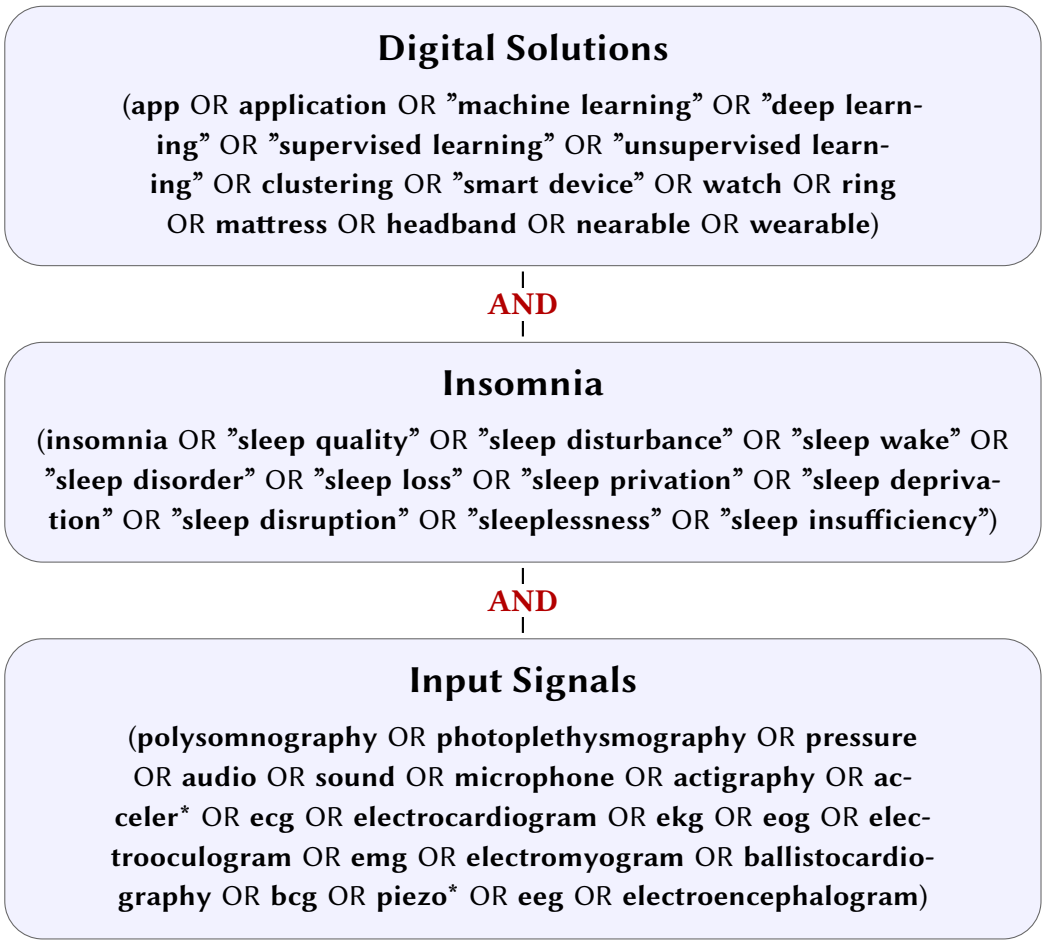


Figure 2.1: Search strategy combining three semantic domains.

All queries, adapted to the specific syntax of each database, are available in Appendix (Table A.1). Queries were performed on February 19, 2024, covering all records available up to that date. Each query results in a list of references corresponding to articles matching the search criteria. The list of references of each database was exported in a Zotero [77] compatible format such as PubMed XML, BibTex, or RIS, as indicated in Appendix (Table A.1). The references were imported into Zotero reference management software [77], and duplicates were removed.

2.3.2 Deduplication

The initial database search yielded a raw corpus of 7,300 records. An automated deduplication process was executed using a customized script within the Zotero reference management software [78] to merge duplicate entries. The dataset was therefore reduced to 4,184 unique references eligible for screening.

Databases	**Interface Platform**	**Nb of References**
Medical Literature Analysis and Retrieval System Online (MEDLINE) (life science journals, and online books)	PubMed (NIH)	1,749
Science Citation Index Social Sciences Citation Index Arts & Humanities Citation Index Conference Proceedings Citation Index Book Citation Index Emerging Sources Citation Index	Web of Science (Clarivate)	1,793
Scopus	Scopus (Elsevier)	3,610
ACM Digital Library	ACM Digital Library	148

Table 2.1: Summary of databases, search queries, and reference counts

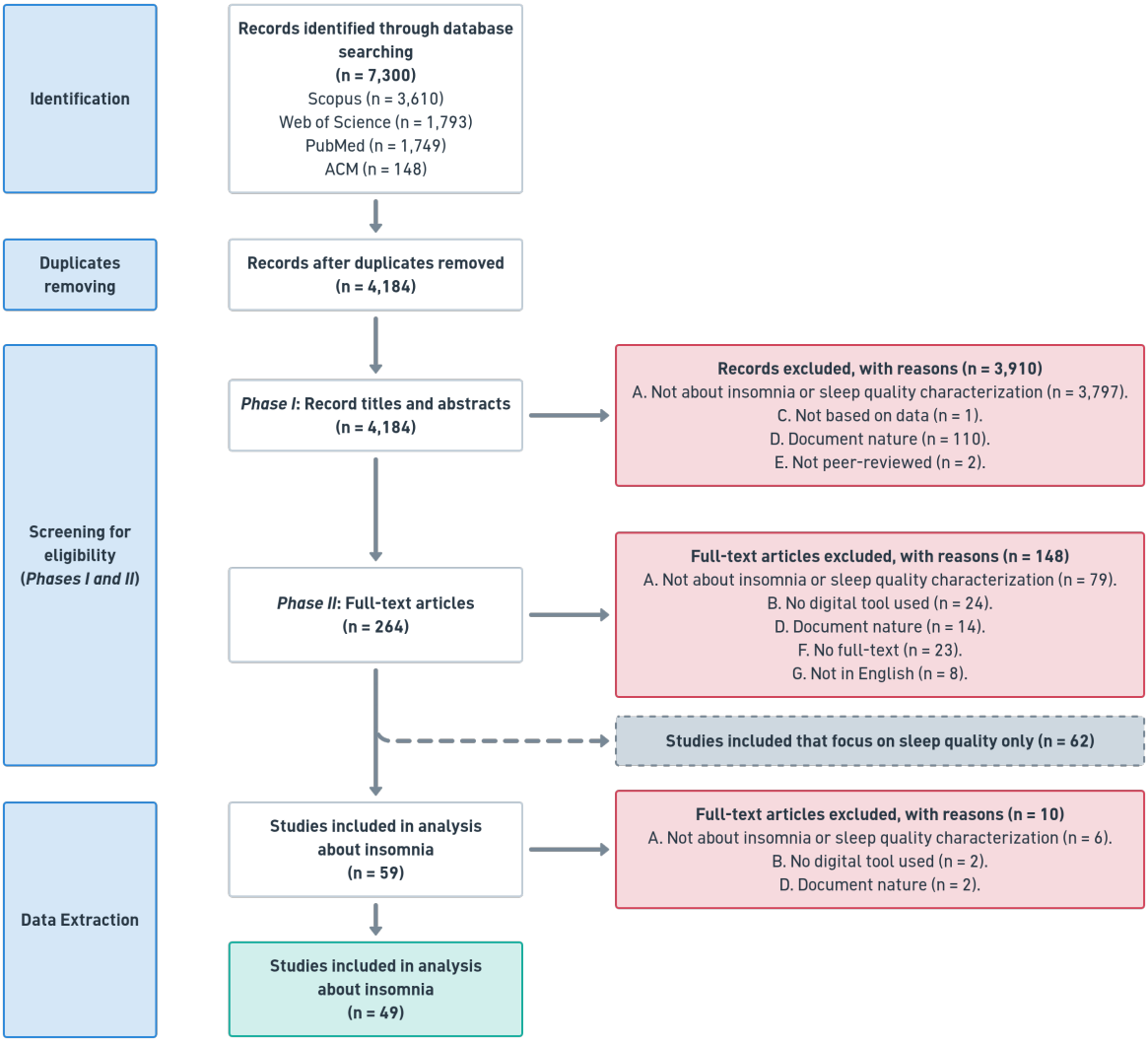


Figure 2.2: PRISMA flow diagram of the systematic review process. The dashed arrow denotes the studies set aside as concerning sleep quality only, kept as a separate subset for potential future analysis rather than counted among the excluded records. The studies retained for the insomnia analysis are those in the green block at the bottom.

2.3.3 Screening process description

The dataset of references underwent a two-phase screening process to determine eligibility for inclusion in the systematic review: the first phase involved title and abstract screening (*Phase I*), followed by a full-text review in the second phase (*Phase II*). The exclusion criteria applied during both phases are summarized in Table 2.2. Eligibility was reassessed during data extraction, and 10 studies were found on closer reading not to meet the inclusion criteria so they were excluded at this stage.

ID	Reason	Description
A	Not about insomnia or sleep quality characterization.	An included reference must bring information about insomnia assessment or detection, or general sleep quality assessment. **Examples of excluded references:** • Research focusing on other pathologies (OSA, schizophrenia, Parkinson, autism, depression, dementia, etc.) • Research focusing on healthy populations only • Research lacking specific focus on sleep quality or insomnia characterization (e.g., circadian rhythm patterns). • References about insomnia treatment. **Examples of included references:** • References about automated insomnia detection through a machine learning model. • References about sleep quality assessment in populations with any sleep disturbances. • References about sleep scoring followed by a use of the results of sleep stages classification to assess sleep quality.
B	No digital tool used.	To be included, the reference must involve the use of any mobile application, sleep monitoring device (actigraphy watch, headband, sleep mattress, etc.) or an algorithm of any nature (signal processing, machine learning, deep learning, etc.).
C	Not based on data.	References that focus on conceptual or philosophical perspective not based on any sort of empirical data were excluded.
D	Document nature.	To focus on primary data, references of the following nature were excluded: systematic reviews, case reports, editorials, letters to the editor, book chapters, comment to a paper, author's response, and preface.
E	Not peer-reviewed.	Papers that have not been peer-reviewed were excluded.
F	No full-text.	References whose full-text were not available were excluded.
G	Not in English.	References whose full-text were not available in English were excluded.

Table 2.2: Criteria for inclusion and exclusion in the systematic review

The screening was performed by a single reviewer due to resource constraints. Another reviewer independently screened a sample of 229 records, 5% of the total, to assess inter-rater reliability.

Exclusion criteria were applied sequentially in the listed order: a record meeting any criterion was excluded on that basis without evaluating the remaining criteria, and a record meeting none was included. During *Phase I*, a conservative strategy retained any record lacking sufficient detail to confirm exclusion, deferring the decision to the full-text review of *Phase II*.

2.3.4 Initial scope definition and scope refinement: insomnia and sleep quality

Because the objective and subjective parameters used to assess insomnia, such as TST, WASO, SE, and SL, also characterize the broader notion of sleep quality [36, 79], the initial search encompassed both the insomnia literature and the general sleep quality literature. A comparative analysis of the two was planned but, due to time constraints, the final synthesis was narrowed to studies of insomnia populations, to focus on distinguishing people with insomnia from those without. The sleep quality records were isolated into a separate subset for potential future analysis.

2.3.5 Data extraction process

The data extraction process involved systematically collecting relevant information from the included studies to facilitate a thorough analysis. We defined the fields to be filled during data extraction, by adapting fields from existing reviews in closely related areas [73, 75, 72].

A standardized data extraction form was developed to ensure consistency. The extracted data items are described in the methods, and the form is available on request.

2.4 Results

2.4.1 Screening results

Phase I screening of 4,184 references against predefined criteria yielded 264 records for full-text review. Phase II full-text review resulted in 59 studies that met all eligibility criteria for the data extraction. During data extraction, an additional 10 studies were excluded upon closer examination, leaving 49 studies for analysis. The number of records excluded at each stage of the screening process, along with the reasons for exclusion, are detailed in the PRISMA flow diagram (Figure 2.2).

2.4.2 Datasets

For comparability across studies, we defined control subjects as those without insomnia as the primary complaint, which may include healthy sleepers or individuals with other sleep or mental health disorders.

2.4.2.1 Publicly available datasets vs. proprietary datasets

Of 49 included studies, 17 (35%) used only publicly available datasets, 32 (65%) used at least one proprietary dataset, and 1 study used both [68]. Some proprietary datasets recur across studies from the same research groups. The publicly available datasets are detailed in Table 2.3. Among the 17 public-only studies, 10 distinct datasets are used and 6 datasets appear in a single study. Table 2.4 provides demographic and clinical characteristics, with the dual-use study counted in both columns, so the 18 public-dataset and 32 proprietary-dataset tallies together span 49 distinct records.

Table 2.3: Overview of publicly available datasets used across selected studies

Dataset Name	Studies (n)	Source / Access URL	References
CAP Sleep Database	8	Physionet	[80, 81, 82, 55, 83, 84, 85, 86]
Holloway et al. (2014)	3	Article Supplementary Materials	[59, 87, 56]
Montreal Archive of Sleep Studies (MASS)	2	Center for Advanced Research in Sleep Medicine	[88, 68]
Insomnia-100	2	Australian New Zealand Clinical Trials Registry	[89, 88]
Circadian Health Database	1	KronoWizard	[90]
Sleep Database Research Center SDRC	1	Article Supplementary Materials	[84]
Sleep Heart Health Study (SHHS)	1	Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC)	[91]
Kermanshah Univ. Medical Sciences	1	Mendeley Data	[68]
Lee et al. (2021)	1	Article Supplementary Materials	[92]
Sleep Health and Lifestyle	1	Kaggle	[93]

Table 2.4: Demographic and clinical characteristics of included datasets

	Publicly Available Datasets ($n = 18$)		**Proprietary Datasets** ($n = 32$)	
Metric	**Mean ± SD**	**Median [Min–Max]**	**Mean ± SD**	**Median [Min–Max]**
Total Participants (n)	119 ± 112	84 [13–447]	409 ± 974	92 [12–4454]
Insomnia (n)	45 ± 55	19 [5–184]	135 ± 179	44 [10–790]
Control[a] Subjects (n)	73 ± 72	66 [0–288]	274 ± 839	30 [0–3985]
Insomnia Prevalence (%)	39 ± 25	40 [8–100]	61 ± 26	53 [11–100]
Control[a] Prevalence (%)	61 ± 25	60 [0–92]	39 ± 26	47 [0–89]

Note: SD = Standard Deviation. [a] Healthy or good sleepers with regard to insomnia. This includes both healthy controls and participants with other sleep disorders that are not insomnia.

2.4.2.2 Assessment of differences across datasets

Besides metrics of dataset size, seven characteristics have been reviewed in the 49 included records:

- the nature and sampling frequency of objective data (e.g. physiological signals) available for each subject
- the monitoring duration per subject
- the temporal distribution of the recorded data, whether recorded at night, during the day, or both
- the inclusion and exclusion criteria used for the population studied, and more specifically the criteria used for subjects with insomnia and for subjects without insomnia
- the presence, nature, and recall of subjective data (e.g. questionnaires, interviews, sleep diaries)
- the study settings (at home vs. in a sleep laboratory)
- recruitment channels

The analysis was performed separately for publicly available datasets and for proprietary datasets, which lets the reader see what data are currently available publicly, without a new collection, and how the two corpora differ.

A. Nature and granularity of objective data

Sampling frequency and sensor availability differ between public and proprietary repositories.

Table 2.5 summarizes sensor availability and typical sampling rates across public and proprietary records. The detail per modality is given in Appendix A.7.

Modality	Public	Proprietary	Typical sampling rate
EEG	✓	✓	100–512 / 200–1000 Hz
EOG, EMG	✓	✓	∼250 Hz (proprietary)
Actigraphy	✓	✓	0.016–0.033 / ≤50 Hz
ECG	4	7	500–600 Hz (proprietary)
oxygen saturation (SpO2)	0	2	N/A
Temperature	1	1	1/600 Hz
Light	1	0	1/30 Hz
Body position	1	1	1/30 Hz

Table 2.5: Objective signal modalities across the 18 public and 32 proprietary dataset records. ✓ marks a modality present in most datasets of that type, electroencephalography being the primary modality and actigraphy the second most prevalent, so their study counts are not enumerated; integers give the number of recording studies for the rarer modalities, with 0 marking absence in this corpus; N/A marks a sampling rate not reported. Sampling rates are given as public / proprietary where they differ. More detail is available in Appendix A.7.

B. Discrepancies in monitoring duration per subject

Monitoring duration per participant ranges from single-session assessments to multi-week longitudinal tracking. Figure 2.3 shows the distribution across public and proprietary dataset records; the full per-study breakdown is given in Appendix A.8.

These discrepancies in monitoring duration may influence the types of insomnia phenotypes that can be detected.

Although reported only once in the studies, the asymmetry in monitoring duration between individuals with insomnia symptoms and controls may introduce a bias in the characterization of insomnia phenotypes. By monitoring healthy controls longer than individuals with insomnia symptoms, the probability of capturing a random night with insomnia symptoms increases. By monitoring individuals with insomnia symptoms for a shorter duration than the controls, some of their symptomatic nights might be missed, therefore artificially reducing the differences between groups.

Importantly, [64] reports the existence of the first night effect, which is the tendency for individuals to sleep poorly in a new environment, for instance in a lab or due to the presence of sensors they are not accustomed to. While a healthy control might look like having insomnia symptoms on the first night of recording, a person with insomnia might sleep more soundly in the lab due to their environmental stressors being removed. Studies limited to a single cross-sectional session of at most 48 hours (29 of 49 included studies, 59%; the cross-sectional bucket of Figure 2.3) might not record sleep representative of the participant's typical pattern, and those capturing a single laboratory night are additionally subject to the first-night effect. This may therefore limit the generalizability of findings.

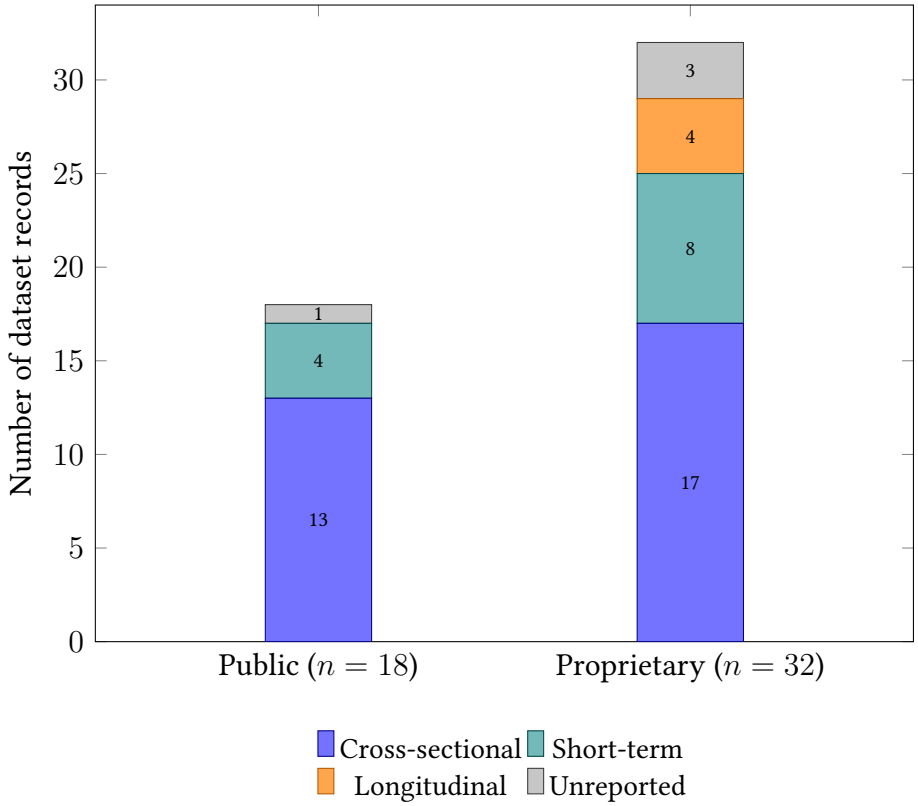


Figure 2.3: Monitoring duration per subject across public and proprietary dataset records. Buckets: cross-sectional ($\leq$48 hours), short-term (3–13 days), longitudinal ($\geq$2 weeks), and unreported. Records basis: the one study using both dataset types [68] is counted in each column. Imaging or punctual studies are counted as cross-sectional, and the single study with asymmetric per-group windows [94] (3–9 nights for individuals with insomnia versus 10–14 for controls) is counted as short-term. The per-study breakdown is given in Appendix A.8.

C. Discrepancies in sleep-wake capture

Beyond duration, the analyses rely on data recorded at different times of the day, which may capture different aspects of insomnia pathophysiology. Studies are classified here by whether the data used for prediction or for the labels are recorded at night, during the day, or both. Figure 2.4 shows the distribution across public and proprietary dataset records, and the studies in each category are listed in Appendix A.9.

Among studies using publicly available datasets, most rely only on night data (78%, $n = 14/18$), while 2 studies (11%, $n = 2/18$) rely only on diurnal data and 2 studies (11%, $n = 2/18$) use both.

Among proprietary datasets, most again rely only on night data (59%, $n = 19/32$), while 9 studies (28%, $n = 9/32$) use both night and diurnal data and 4 studies (13%, $n = 4/32$) rely only on diurnal data, likely targeting biomarkers of daytime activity or hyperarousal rather than sleep architecture.

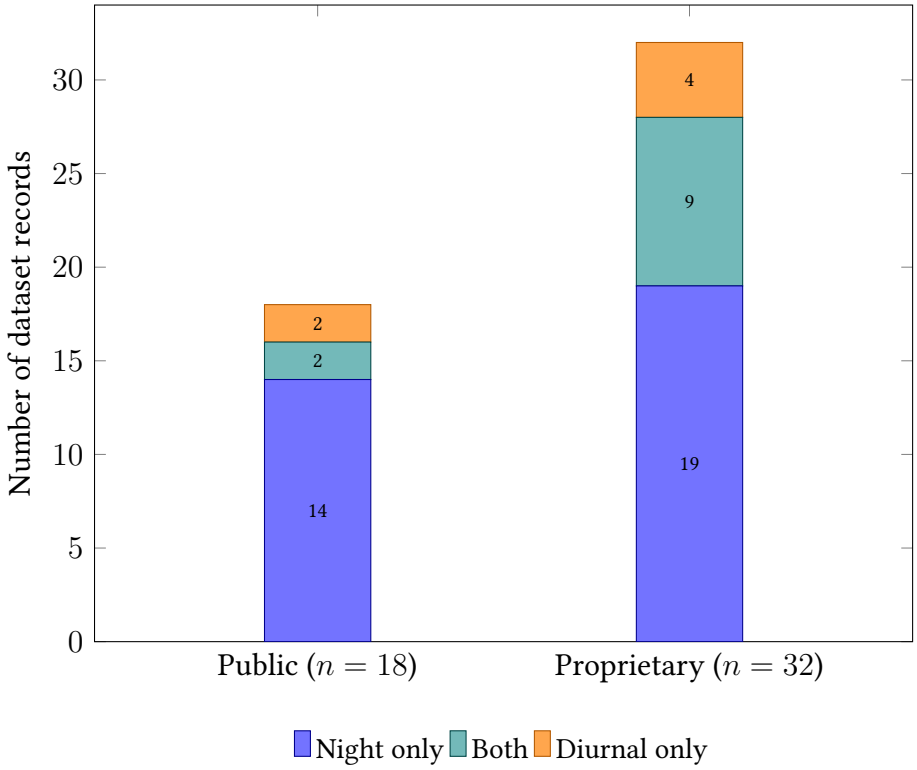


Figure 2.4: Night and diurnal data capture across the 18 public and 32 proprietary dataset records, classified by whether the data used for prediction or for the labels are recorded at night, during the day (diurnal), or both. The study [68] is counted in public and proprietary categories. The studies in each category are listed in Appendix A.9.

This suggests that a substantial proportion of proprietary studies (41%) use diurnal data, compared with only 22% of those using publicly available datasets. This may reflect a gap in the current understanding of insomnia pathophysiology and its relationship with daytime functioning, which could be an avenue for future research.

D. Inclusion and exclusion criteria

The review of inclusion and exclusion criteria across studies reveals wide heterogeneity in the characteristics of insomnia and control groups, which may impact the generalizability and comparability of findings.

Both groups share exclusion criteria across the same broad domains: medical and psychiatric comorbidities, sleep comorbidities, pharmacological exposures, lifestyle and occupational factors, physical health conditions, and technical constraints. The specific criteria and quantitative thresholds within each domain differ markedly across studies.

The six recurring exclusion domains and representative criteria are summarized in Table 2.6. The full description of all exclusion criteria is provided in Appendix A.3.

Criteria specific to control-group subjects

Table 2.6: Most representative exclusion criteria across the included studies. Full per-study detail is provided in Appendix A.3.

Domain	Main criteria	References
Medical and psychiatric comorbidities	Neurological or psychiatric history, cognitive impairment, traumatic brain injury; verified by structured clinical interview or psychometric screening (Hospital Anxiety and Depression Scale, Beck inventories).	[95, 44, 96, 92, 60, 59, 87, 90, 97, 98, 99, 45, 100]
Sleep comorbidities	Obstructive sleep apnea (Apnea-Hypopnea Index $>$ 5–10), periodic limb movement disorder, restless legs syndrome, narcolepsy.	[90, 92, 99, 98, 91, 101, 94, 96, 102, 44, 45, 103, 97, 104]
Pharmacological and substance exposures	Sedatives, hypnotics, psychoactive medication, illicit substances, high caffeine intake; washout periods from five days to two months.	[95, 44, 101, 91, 45, 98, 105, 99, 97, 96, 100, 92, 106, 55]
Physical health conditions	Cardiovascular, respiratory, or metabolic disease; Body Mass Index (BMI) of 30 kg/m^2 or greater; pregnancy, lactation, or breastfeeding.	[91, 96, 102, 94, 99, 98, 105, 95, 92]
Lifestyle, occupational, and circadian factors	Shift work, transmeridian travel, irregular sleep-wake rhythms, ongoing behavioral therapy or counselling.	[99, 92, 101, 107, 45, 96, 105]
Technical and operational constraints	Magnetic resonance imaging or transcranial magnetic stimulation contraindications, smartphone requirements, and minimum data-quality thresholds.	[95, 92, 105, 108, 107, 109, 110, 56, 60, 91]

The classification of healthy control subjects relies on heterogeneous criteria. Some studies use subjective self-reporting of unimpaired sleep [59], while others apply standardized psychometric thresholds such as a PSQI score below 5 and an ISI score below 8 [100, 95].

Excluding individuals presenting a body mass index of 30 or greater, habitual snoring, or mild sub-clinical anxiety generates highly restricted control cohorts [94, 96]. While such stringent screening serves the purpose of maximizing group separability during model development, it introduces spectrum bias. Classifiers trained to distinguish clinical insomnia from strictly screened normative sleep profiles may inflate apparent sensitivity and specificity during development, yet degrade in discriminative performance when deployed on heterogeneous populations where sub-clinical sleep disturbances, elevated weight, or snoring co-occur without clinical insomnia. In at least one study, the converse situation is observed: defining control groups through subjective self-declaration while subjecting the insomnia group to PSG screening introduces differential misclassification bias [59]. Subjectively defined normative sleepers may harbor asymptomatic obstructive sleep apnea, intro-

ducing confounding variance into the training set and reducing predictive accuracy. The disparate rigor in control group definitions might therefore degrade classifier robustness and generalizability to real-world populations containing sub-clinical sleep disturbances and common comorbidities.

Criteria specific to insomnia-group subjects

In insomnia subjects, the inclusion criteria often require a clinical diagnosis based on standardized criteria such as DSM-5 or ICSD-3, but the specific thresholds for comorbidities, symptom severity, duration, and associated distress vary widely across studies, producing heterogeneous insomnia phenotypes.

A large portion of the studies using proprietary datasets failed to report the inclusion and exclusion criteria for the studied population, representing 41% ($n = 13$) for the healthy group and 19% ($n = 6$) for the insomnia group.

The filtering applied in insomnia research usually tries to isolate insomnia as the primary complaint from insomnia symptoms due to clinical, pharmacological, and lifestyle confounders. However, the dimensions along which criteria diverge most, namely diagnostic classification systems, comorbidity thresholds, and pharmacological washout periods, produce cohorts that differ not only in size but in the clinical profile of the insomnia phenotype under study. While some of this heterogeneity reflects legitimately different research questions, the resulting variation restricts the reproducibility and cross-study comparability of machine learning models trained on these specific cohorts. Furthermore, a substantial proportion of investigations using proprietary datasets fail to report their inclusion and exclusion parameters entirely, therefore undermining methodological transparency. Future investigations would benefit from adopting standardized clinical definitions and reporting guidelines to improve the validity, replicability, and generalization of automated sleep classification systems.

E. Collection of subjective data

Public datasets contain markedly less subjective data than proprietary datasets, as 78% of the studies ($n = 14$) derived from four main publicly available datasets contain zero subjective data. Only four studies integrate subjective assessments [91, 90, 92, 93].

Among proprietary datasets, 41% of the studies ($n = 13$) did not use subjective data to understand insomnia.

When subjective data are collected, the instruments span insomnia-severity, sleep-quality, sleepiness, sleep-diary and ecological-momentary-assessment, mood, and diagnostic categories, summarized in Table 2.7. The detailed description is given in Appendix A.4.

Subjective data collection in insomnia research suffers from methodological fragmentation, with 78% of public dataset studies and 41% of proprietary dataset studies omitting subjective data entirely. When subjective metrics are collected, the contrast between their near-total absence in public repositories and the abundance of diverse questionnaires in proprietary datasets restricts cross-study validation. For studies aiming at clinical translation, aligning objective physiological findings with standardized patient-reported outcomes

Table 2.7: Subjective instruments used across the included studies, grouped by category. Detail is available in Appendix A.4.

Category	Instruments	References
Diagnostic criteria and clinical histories	DSM-5, ICSD-3, Research Diagnostic Criteria; unstructured sleep and medical histories	[110, 44, 109, 102, 107, 90]
Insomnia severity	Insomnia Severity Index, Athens Insomnia Scale	[110, 44, 60, 105, 67, 111]
Sleep quality	Pittsburgh Sleep Quality Index, mini sleep questionnaire	[92, 44, 105, 108, 101]
Sleepiness and fatigue	Epworth Sleepiness Scale, Fatigue Severity Scale, visual analog scale	[92, 44, 112, 113, 105, 46, 45]
Sleep diaries and ecological momentary assessment	Pittsburgh Sleep Diary, Evening-Morning Sleep Questionnaire, ecological momentary assessment, subjective sleep-parameter estimates	[110, 101, 112, 94, 106, 98, 114, 97, 107, 91, 113, 46]
Depression and anxiety	Beck Depression and Anxiety Inventories, Zung Depression Inventory, PHQ-9, Cattell and Scheier Anxiety Scale, non-standardized mood metrics	[44, 109, 108, 93, 92]
Broad psychiatric and general health	Minnesota Multiphasic Personality Inventory, Multiple Affect Adjective Check List, Profile of Mood States, Cornell Medical Inventory, SF-36	[109, 45, 111]
Dysfunctional sleep beliefs	Dysfunctional Beliefs and Attitudes about Sleep Scale-16	[92]

would strengthen the reproducibility and external validity of automated classification systems.

Subjective data are also seldom collected repeatedly over time: most studies that include them sample once, whereas ten capture them recurrently through multi-day sleep diaries, daily questionnaires, or ecological momentary assessment [110, 101, 112, 94, 106, 98, 114, 97, 107, 111], all drawn from proprietary datasets. Recurrent subjective monitoring more closely reflects the night-to-night fluctuation that defines insomnia, so its scarcity is a notable gap.

F. Recording environments: laboratory vs. home settings

The physical recording environment for insomnia sleep data varies between controlled sleep laboratories and naturalistic home settings. Among public datasets, researchers recorded data in a sleep laboratory in $n = 12$ (67%) studies and at home in $n = 4$ (22%) studies, while $n = 1$ (6%) study records no nocturnal sleep, so no setting applies, and $n = 1$ (6%) study does not indicate the recording environment. Within proprietary datasets, $n = 18/32$ (56%) studies use sleep laboratories, though the specific environment remains ambiguous in [115]. Furthermore, $n = 10/32$ (31%) proprietary studies use home recordings, with the environmental classification remaining ambiguous in [101]. Finally, for the remaining proprietary studies the sleep environment does not apply or is unstated: $n = 2/32$ (6%) record no nocturnal sleep, so no laboratory-or-home setting applies, and $n = 2/32$ (6%) do not report the environment.

While controlled laboratory environments remain the dominant setting for acquiring objective data on insomnia disorder, naturalistic home monitoring constitutes a substantial portion of the literature. This distribution reflects a clinical trade-off between recording modalities. Sleep laboratories provide high environmental control to isolate physiological signals and minimize external artifacts. Conversely, home settings offer necessary ecological validity to capture typical sleep behaviors aligned with the daily routines of the individuals.

G. Recruitment channels

Recruitment channels for insomnia studies rely heavily on clinical environments, though researchers frequently supplement these populations with public advertising or institutional outreach. Public-dataset participants come primarily from hospital sleep disorder centers, while proprietary studies draw on university-affiliated sleep centers and hospital clinics, supplemented by advertising, radio and print media, community outreach, national registries, and targeted occupational populations. The details are available in Appendix A.5. Nonetheless, 21% of the studies using proprietary datasets ($n = 7$) do not indicate their recruitment channels.

H. Alternative recording tools for insomnia phenotyping

Among the 49 included studies, wrist actigraphy and polysomnography are the dominant recording modalities. Three studies use tools that extend beyond these two standard methods, summarized in Table 2.8. A more detailed per-device description is given in Appendix A.6.

These three tools capture signals or contexts not available from standard wrist actigraphy alone: multichannel circadian monitoring adds temperature and ambient light channels, chest-worn ECG enables cardiopulmonary coupling analysis, and passive smartphone tracking provides an unobtrusive proxy for circadian rhythms that does not require a clinical device. Each tool appears in a single study within this corpus, and no second study replicates or independently validates any of the three approaches for insomnia characteri-

Table 2.8: Recording tools allowing insomnia phenotyping beyond standard actigraphy and polysomnography. Full per-device descriptions are given in Appendix A.6.

Tool	Signals captured and output	Reference
Kronowise ambulatory circadian monitoring	Skin temperature, motor activity, body position, ambient light; outputs an estimated sleep probability and circadian rhythm indices.	[90]
M1-SleepImage chest wearable	Continuous ECG at 600 Hz, body position, snoring; cardiopulmonary coupling metrics, sleep spectrogram, actigraphic total sleep time and wake after sleep onset.	[94]
Rhythm smartphone application	Screen on/off events, notifications, application usage; six sleep indicators and three circadian rhythm metrics.	[108]

zation.

I. Conclusion on datasets

Most included studies rely on proprietary datasets, representing 65% of the literature, while publicly available repositories remain fragmented across ten different datasets, six of which appear in a single study, leaving the field without a consensus benchmark for reproducible evaluation. The signal modalities available differ between the two categories: pulse oximetry appears only in proprietary datasets and ambient light only in public repositories, while ECG, body temperature, and body position appear in both, so that no single dataset category captures the full range of physiological dimensions relevant to insomnia phenotyping. The temporal scope of data acquisition further constrains insomnia characterization, as 59% of studies rely on a single cross-sectional session of at most 48 hours, which precludes the assessment of night-to-night variability and the symptom frequency criteria required by DSM-5. In addition, 78% of public-dataset studies and 59% of proprietary-dataset studies rely only on night data, leaving daytime functioning unexamined despite its relevance to insomnia diagnosis. Population definitions diverge across diagnostic classification systems, comorbidity thresholds, and pharmacological washout periods, producing cohorts with distinct insomnia phenotypes, while overly strict control group screening introduces spectrum bias that may inflate classifier performance during development relative to deployment on real-world populations. Subjective data collection remains largely absent, with 78% of public-dataset studies and 41% of proprietary-dataset studies integrating no patient-reported outcomes, despite insomnia being partly defined by subjective complaint. Recording environments span controlled laboratories and naturalistic home settings, while several studies use divergent recruitment channels for insomnia and control groups, introducing potential selection bias, and reporting transparency remains poor, as 41% of proprietary studies do not disclose inclusion criteria for the healthy group and 21% omit recruitment methods. Collectively, benchmark fragmentation, incomplete signal coverage,

insufficient temporal resolution, heterogeneous population definitions, and absent subjective data produce non-comparable insomnia phenotypes and restrict the reproducibility and clinical translation of automated classification results. The feature extraction methodologies and classification models reviewed in the subsequent section should therefore be evaluated in light of these data-level constraints.

2.4.3 Methods

2.4.3.1 Input data

A. Data preprocessing

A.a. Missing data

Among the subset of studies that document missing data procedures, missing data occurs in two contexts: incomplete raw data time series, such as actigraphy, and missing entire recording nights. Reported prevalence is minimal, with one study noting approximately 1% missing values in actigraphy time series.

The handling strategies used vary across studies. Six studies adopt deletion approaches, removing subjects or individual nights with missing data [98, 113, 89, 91, 56, 97]. One study excludes subjects from analyses for which their data are incomplete, but retains them in analyses where complete data are available [94]. Two studies implement imputation methods: one applies a moving median filter with a 30-minute sliding window to fill gaps in actigraphy time series [87]. Another study uses multiple imputation via random forest algorithms.

This heterogeneity in missing data handling is consistent with the absence of standardized procedures for the field. Deletion-based approaches risk sample bias and reduced statistical power, whereas imputation strategies introduce model-dependent assumptions. The limited documentation of missing data frequency and handling methods across most studies prevents formal assessment of the resulting impact on model generalizability and comparability across studies.

A.b. Preprocessing steps

Signal preprocessing encompasses operations that directly condition raw signals prior to feature extraction, including filtering, artifact removal, normalization, temporal discretization, and resampling. Signal preprocessing complexity varies substantially across input modalities. Sixteen studies do not report any data preprocessing prior to analysis [80, 110, 45, 101, 94, 106, 67, 113, 116, 97, 91, 56, 111, 115, 102, 114]. Among studies using EEG signals, multi-stage pipelines are common, whereas actigraphy-based and questionnaire-based studies typically apply minimal or no data preparation steps.

The specific operations applied by each study are detailed in Appendix A.13.

Conclusion

Preprocessing approaches across included studies exhibit substantial heterogeneity, with no standardized protocol emerging as dominant. Electroencephalography studies use multi-stage pipelines incorporating bandpass filtering, artifact rejection, fixed-duration segmentation, normalization, and occasional resampling, anchored to American Academy of Sleep Medicine (AASM) scoring conventions and frequencies derived from established sleep neuroscience. Actigraphy-based investigations apply minimal raw signal processing, delegating substantially to feature-level transformations, with isolated use of logarithmic transformation, discretization, and concatenation of daily logs in sliding windows. Neuroimaging modalities use domain-specific preprocessing workflows with specialized software tools, and studies relying exclusively on questionnaire-based measures require no signal processing.

B. Feature extraction

B.a. Cross-sectional demographic and clinical data

In addition to continuous physiological recordings, six studies incorporate cross-sectional demographic and clinical baseline measurements to contextualize the main sleep or activity data. Age and sex represent the most frequently integrated baseline features [106, 105, 104, 46, 93, 112]. Researchers frequently include general anthropometric indicators such as BMI [106, 104, 46, 93, 112], alongside more specific physical measurements including neck and waist circumference [104] and waist-hip ratio [46].

Broader socioeconomic and demographic descriptors encompassing education, race, ethnicity [106, 46], and occupation [93] occasionally are used as inputs to capture demographic diversity and health disparities. Furthermore, some studies explicitly integrate underlying clinical health baselines to guide the models, incorporating factors such as systolic and diastolic blood pressure [93] and established psychiatric diagnoses like depression [112].

B.b. Traditional sleep and activity parameters

Traditional sleep and activity parameters constitute the most frequently used features across the included studies, deriving directly from standard clinical evaluations and providing direct clinical interpretability through alignment with conventional clinical sleep assessment. As detailed in Table 2.9, these parameters are categorized into sleep architecture and timings, awakenings and disruption indices, motor activity profiles, and specific physiological events. While both actigraphy and PSG capture overlapping single-night sleep-wake parameters, each modality also generates unique, method-specific features. For instance, actigraphy provides non-parametric circadian rhythm features, namely Interdaily Stability (IS) and Intradaily Variability (IV), to capture multi-day behavioral rhythms, whereas PSG provides cardiovascular, respiratory, and neuromuscular event indices.

Beyond the core parameters of Table 2.9, studies derive variants specific to each recording method. Actigraphy yields more detailed measures such as daily step counts, nap du-

Table 2.9: Traditional sleep and activity parameters used across insomnia datasets

Feature category	Parameters	Records using actigraphy	Records using PSG
Sleep architecture & Timings	Total Sleep Time (TST)	[110, 99, 59, 94, 87, 108, 107, 117, 93]	[106, 91, 112, 97, 117, 98, 113, 103, 46, 114]
	Sleep Latency (SL) / Latency to Persistent Sleep (LPS)	[99, 101, 117]	[44, 45, 112, 97, 117, 98, 113, 103, 109, 46, 114]
	Sleep Efficiency (SE)	[99, 101, 94, 117]	[44, 112, 117, 113, 46, 114]
	Time in Bed (TIB)	[107]	[112, 97]
	Stage durations and/or proportions	[117]	[44, 112, 97, 117, 113, 103, 109, 114]
	REM Latency / Sleep Onset REM	[117]	[117, 113, 109, 46, 114, 112]
	Sleep Onset, Wake Time, Mid-Sleep Time	[90, 108]	[114]
	Sleep Onset, Wake Time	[108]	[114]
Awakenings & Disruptions	Wake After Sleep Onset (WASO)	[99, 59, 94, 87, 108, 107, 117]	[112, 117, 98, 113, 103, 114]
	Number of Awakenings / Arousals	[94, 108]	[103]
Motor & Circadian activity	IS	[90, 87, 108]	-
	IV	[90, 87, 108, 56]	-
	Relative Amplitude (RA), L5, M10, CFI	[90, 108]	-
	Body Position	[94]	-
Physiological events (PSG)	Heart Rate (HR)	-	[44, 45, 46, 93]
	AHI & Respiratory Disturbance	-	[112, 46]
	Oxygen Desaturation (SpO2, DI)	-	[104, 112, 46, 113]
	Periodic Leg/Limb Movement Index	-	[112, 46]

rations, the ratio of sleep to wake, and statistics of movement intensity, while PSG yields additional temporal and structural markers such as arousal indices, indices of transitions between sleep stages, the timing of sleep fragmentation across the night, and discrete event markers including snoring episodes and body movements. A few studies derive parameters to control for confounders such as the first-night effect. These variants are detailed in Appendix A.12.

B.c. Mathematical data transformations

Where traditional parameters provide macroscopic summaries of the sleep-wake cycle, mathematical data transformations characterize the fine structure of physiological signals, extracting continuous features related to spectral power, statistical distribution, and non-linear complexity.

For actigraphy, these methods typically involve treating the continuous movement data as a time series, allowing for the extraction of entropy indices, fractal dimensions, and complex scaling exponents. These metrics measure the predictability, self-similarity, and irregularity of nocturnal motor activity. For PSG recordings, mathematical transformations quantify the spectral power of specific EEG frequency bands, including delta, alpha, and gamma, across different brain regions and sleep stages, as well as across different physiological systems such as ECG and respiration.

Table 2.10 synthesizes these computational features across both modalities, providing the underlying clinical or physiological intuition for why these mathematical models are applied to sleep data.

Beyond the transformations of Table 2.10, a subset of studies apply methods based on models or on signal decomposition: unsupervised autoencoders that compress daily actigraphy features, wavelet decomposition of EEG and ECG signals concentrated within a single research group, spectral power computed with the Fourier transform across canonical frequency bands, and extraction of EEG graphoelements such as slow waves and sleep spindles together with the slope of the aperiodic spectrum. A further set of studies model functional brain connectivity and non-linear EEG dynamics. These methods are detailed in Appendix A.12.

B.d. Functional magnetic resonance imaging and spectroscopy

Beyond signal processing of sleep-wake recordings, four of the 49 included records used neuroimaging: three applied functional MRI to extract connectivity and regional metrics, and one quantified neurochemical markers through spectroscopy. These approaches capture static neural properties complementary to physiological sleep-wake phenotypes, such as structural abnormalities and functional connectivity patterns not directly measurable by actigraphy and PSG.

The specific cortical and subcortical parcellations, regional metrics, and neurochemical measures used by these four studies are detailed in Appendix A.12.

Table 2.10: Mathematical data transformations and their clinical intuition

Feature Category	Transformation	Intuition	Modalities & References
Statistical Moments & Distributions	Mean, SD, Variability Analysis	Elevated variability in nocturnal actigraphy reflects increased motor restlessness and fragmented sleep characteristic of insomnia.	Actigraphy [99, 87, 116]
	Complementary Cumulative Distribution Function (CCDF)	Quantifies the probability that nocturnal motor activity exceeds a given intensity level, capturing the frequency of sleep-disruptive movement episodes.	Actigraphy [87]
	Arousal Statistical Distributions	Characterizes sleep fragmentation severity through the full statistical distribution of arousal events, which are typically elevated in insomnia.	PSG [112]
Entropy & Information	Sample Entropy (SampEn)	Quantifies the regularity of nocturnal signals, where altered predictability in motor or EEG activity reflects the disrupted sleep microstructure of insomnia.	Actigraphy [99, 59], PSG [115]
	Shannon Entropy (ShE)	Measures the information content of the activity signal, where higher entropy indicates more disordered and less consolidated nocturnal rest.	Actigraphy [87]
Dynamics & Complexity	Poincaré Map Parameters (SD1, SD2, ratio)	Quantifies short-term and long-term variability of successive activity values, robust to the noisy motor artifacts common in nocturnal actigraphy.	Actigraphy [99, 59]
	Detrended Fluctuation Analysis (DFA) & Power Law Index β	Detects long-range temporal correlations in nocturnal signals, where deviation from healthy scaling exponents indicates altered sleep regulation across the night.	Actigraphy [116, 87], PSG [115]
	Magnitude and Sign Analysis (MSA)	Decomposes activity fluctuations into magnitude and direction components to detect nonlinear asymmetries in nocturnal motor patterns that differ between insomnia and healthy sleep.	Actigraphy [116]
	Recurrence Quantitative Analysis (RQA)	Analyzes recurring patterns in the phase-space trajectory of nocturnal activity, where reduced determinism reflects the irregular, non-periodic sleep fragmentation of insomnia.	Actigraphy [116]
	Higuchi Fractal Dimension (HFD)	Estimates the fractal dimension of the activity signal, where higher complexity reflects the irregular, multi-scale motor patterns associated with disrupted sleep.	Actigraphy [87]
	Hjorth Parameters (Activity, Mobility, Complexity)	Characterizes the oscillatory properties of the EEG signal through variance, mean frequency, and bandwidth, detecting altered neural dynamics in insomnia without requiring Fourier stationarity assumptions.	PSG [118, 82, 83, 53]
Spectral & Rhythm Analysis	EEG Band Power	Quantifies spectral energy in frequency bands tied to distinct sleep processes, enabling detection of altered delta or beta activity associated with insomnia hyperarousal.	PSG [112, 118, 53]
	Power Spectral Analysis (PSA) parameter	Extracts the power spectrum of actigraphy to identify self-similar scaling properties that differ between consolidated healthy sleep and fragmented insomnia patterns.	Actigraphy [116]
	Cosinor & Singular Spectrum Analysis	Models the 24-hour rest-activity cycle to detect disruptions in circadian amplitude, phase, or stability associated with insomnia.	Actigraphy [56]
Cross-Signal Coupling	Cardiopulmonary Coupling Metrics	Quantifies the synchronization between cardiac and respiratory rhythms during sleep, where reduced high-frequency coupling reflects the autonomic hyperarousal associated with insomnia.	PSG [94]

B.e. Subjective data including questionnaires

In addition to objective physiological recordings, research studies investigating insomnia biomarkers systematically incorporate subjective self-report measures to capture individuals' perceptions of sleep quality, daytime functioning, and psychological distress. These measurements are a necessary complement to objective data, as they quantify the subjective symptomatology that defines insomnia as a clinical disorder [36]. The literature reveals marked heterogeneity in questionnaire selection, reflecting the absence of standardization around which measures provide optimal diagnostic utility for insomnia characterization.

The instruments considered here are those incorporated as input features for automated insomnia characterization, excluding instruments used solely for descriptive purposes or diagnostic group assignment.

Across these studies, the questionnaires used as input features span seven categories: scales specific to insomnia such as the Pittsburgh Sleep Quality Index, the Insomnia Severity Index, and the Athens Insomnia Scale; inventories of psychiatric and mood symptoms; scales of daytime sleepiness and fatigue such as the Epworth Sleepiness Scale; screens for substance use; sleep parameters reported by the participants through single ratings and diaries kept over several nights; ecological momentary assessment; and measures of quality of life. The full list of instruments, with the studies that use each one, is given in Appendix A.12.

How and how often the questionnaires were administered varies widely: single-occasion baseline questionnaires [106, 44, 112, 46], sleep diaries spanning seven to fourteen nights [106, 114, 107], and ecological momentary assessment with multiple daily prompts [107]. This variation in measurement frequency reflects differing study designs, from cross-sectional phenotyping to mechanistic investigation of within-person variability.

B.f. End-to-end models

In contrast to the feature engineering approaches described above, a minority of studies used end-to-end deep learning models that learn representations directly from raw physiological signals without explicit feature extraction. Only 3 of the 49 included studies, representing 6%, adopted this strategy [88, 68, 80], and all three exclusively operated on EEG signals. Low adoption may reflect limited data availability, computational requirements, or the interpretability demands of clinical applications.

C. Feature selection and dimensionality reduction

Once features are extracted, some studies further reduce the dimensionality of the feature vector to counter overfitting, to mitigate the curse of dimensionality in distance-based clustering, to enforce statistical independence required by downstream methods, or to gain interpretability. Ten of the forty-nine included studies report a post-extraction dimensionality-reduction step; the remaining studies feed the extracted feature vector directly to the classifier or clustering method. These ten studies are surveyed below by method family: univariate relevance filters, multivariate redundancy filters, wrapper methods, embedded methods,

and unsupervised projection.

C.a. Univariate relevance filters

Filter methods rank features through statistical scores computed between each feature and the class label, independently of any learning algorithm. An actigraphy study applied information gain, analysis of variance, and Chi-squared testing to a set of candidate circadian-rhythm indices, retaining those with discriminative potential for sleep-disorder classification [90]. Another study combining physiological and questionnaire inputs compared Minimum Redundancy Maximum Relevance (mRMR), Chi-squared, analysis of variance, and the Kruskal-Wallis test, selecting mRMR as the method yielding the highest downstream classification accuracy [93]. Among these four methods, mRMR differs in combining a relevance score with a redundancy penalty, whereas the remaining three rank features independently of one another.

C.b. Multivariate redundancy filters

Multivariate filters score features jointly, targeting redundancy among correlated features rather than relevance to the class label. One actigraphy study applied the Goldbricker procedure from the `networktools` R package to select a minimally-correlated subset of features prior to Latent Class Analysis (LCA), which assumes the features within each latent class to be pairwise independent, so a minimally-correlated subset is required; the authors preferred Goldbricker over Principal Component Analysis (PCA) for this purpose [110]. One PSG-based study started from an initial set of 152 PSG, clinical, and questionnaire features and removed 25 of them through a pair-wise Pearson correlation cutoff at $|r| > 0.9$ [112]. A second PSG-based study applied Infinite Feature Selection (Inf-FS), an unsupervised filter that balances correlation against variance to score features, reducing 54 spectral features to 6 before subsequent projection [89]. An actigraphy study additionally removed time-in-bed *a priori* because of its high correlation with total sleep time, narrowing the daily-feature set before unsupervised projection [60].

C.c. Wrapper methods

Wrapper methods evaluate feature subsets through the predictive performance of a target learning algorithm. The PSG-based study using the pair-wise correlation cutoff subsequently applied Recursive Feature Elimination (RFE) with logistic regression and five-fold CV, removing 55 features and yielding a final subset of 72 [112]. No other included study reports a wrapper step as its primary selection mechanism.

C.d. Embedded methods

Embedded methods select features during model fitting through regularization that penalizes model complexity by shrinking or zeroing coefficients. One neuroimaging study applied Least Absolute Shrinkage and Selection Operator (LASSO) logistic regression with L1 regularization, the penalty parameter λ selected by LOO CV, retaining features surviving

the L1 penalty as input to a SVM classifier [92]. This embedded step follows the region-of-interest PCA described below. No other included study reports an embedded selection step distinct from the inductive bias of its main classifier.

C.e. Unsupervised projection

Five studies project the feature vector onto a lower-dimensional subspace by linear methods and one by a non-linear method. PCA extracts orthogonal components that maximize explained variance, whereas factor analysis models the observed variables as a smaller set of latent common factors. Two cluster-analysis studies use factor analysis, one narrowing 142 candidate variables to 26 inputs with a literature search and clinical judgement [109] and one reducing 38 sleep-history variables to 15 composite factors [102]. Three studies use PCA: a PSG phenotyping study keeps 2 components of the 6 Inf-FS-retained spectral features [89], and two neuroimaging studies reduce their high-dimensional feature sets to a few components for insomnia-subtype classification [92, 95]. For the non-linear case, one actigraphy study trains a convolutional Autoencoder (AE) that compresses an eight-day by twelve-feature activity tensor into 15 latent variables, then applies t-distributed Stochastic Neighbor Embedding (t-SNE) to embed them in two dimensions for k-means clustering [60].

C.f. Conclusion

Within this corpus, explicit dimensionality reduction concentrates in two study types: multimodal neuroimaging studies with feature counts comparable to or exceeding subject counts, and phenotyping or clustering studies in which distance-based or probabilistic cluster assignment degrades under high dimensionality. Single-channel EEG classification studies in this corpus feed the extracted feature vector directly to the classifier without an explicit reduction step, relying on the classifier's built-in regularization to manage dimensionality. Stated rationales can be grouped into three patterns: overfitting control, interpretability of retained features, and statistical-independence requirements of downstream methods such as LCA. Methodological comparisons between reduction strategies are rare: one PSG-based phenotyping study benchmarks its Inf-FS and PCA pipeline against a multi-cluster feature selection alternative [89], and one questionnaire-based study contrasts mRMR with three univariate alternatives [93], with no further study contrasting reduction strategies within its own experimental design. Clinical judgement informs the choice of the initial feature pool, and a purely statistical criterion drives the subsequent reduction step. One cluster-analysis investigation is distinctive in integrating clinical judgement directly into the reduction step itself, alongside factor analysis and a literature search [109].

2.4.3.2 Models and outputs

The 49 included studies are organized by their output type. Figure 2.5 shows their distribution across the four types, and the algorithms producing each output are detailed per study in Appendix A.11. Four output types, each study assigned to a single category based on its primary output, are identified: binary classification in 17 studies, multi-class or multi-task classification in 10, cluster-based phenotypic stratification in 13, and quantitative statis-

tical or signal-processing measures in 9.[1] For the purposes of this review, deep learning designates studies in which a convolutional, recurrent, or generative deep architecture is the sole classifier; studies that apply such architectures alongside conventional machine learning classifiers are instead classified as combining machine learning and deep learning.

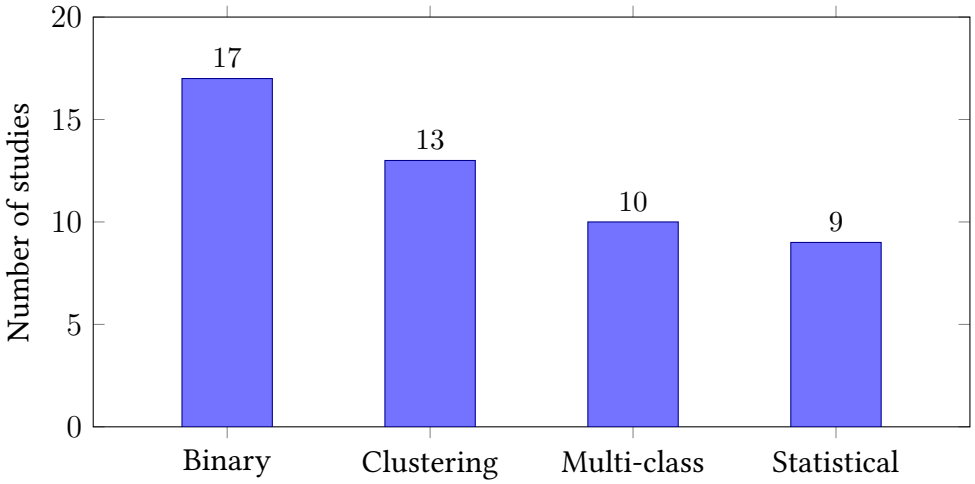


Figure 2.5: Distribution of the 49 included studies by primary output type. Each study is assigned to a single category based on its primary output. Per-study models and algorithms are detailed in Appendix A.11.

Binary classification distinguishing a control group from a single insomnia category is the most common output structure, appearing in 17 studies. Ten studies extend beyond binary classification to discriminate among multiple sleep pathologies or insomnia subtypes, with class counts ranging from three to eight. Cluster-based phenotypic stratification is used by 13 studies as an unsupervised approach to insomnia subtyping that does not require pre-defined diagnostic labels, with cluster counts ranging from two to 14 and three-cluster solutions reported in five of the 13 studies, the most frequent outcome across the sample. The remaining nine studies produce quantitative statistical or signal-processing measures rather than discrete classifications or cluster assignments, seven relying on statistical inference and two applying signal-processing methods without subsequent classification.

Although most of the 49 studies produce a single prediction or cluster assignment per subject, two exceptions operate at different granularities: Angelova et al. generate nightly predictions aggregated to subject level via a frequency threshold [59], and Park et al. produce temporal sequences of cluster assignments across 35 eight-day windows subsequently collapsed to individual-level phenotypes [60]. This variation across output granularities, combined with the absence of standardized output mapping, prevents formal cross-study comparison of probabilistic estimates with hard labels, cluster assignments with diagnostic categories, and signal-processing measures with categorical phenotypes.

[1]Output type is assigned by primary output structure, which differs from the validation taxonomy of the Validation-method subsubsection: Tonetti et al. [101] is counted here under binary classification but among the statistical, no-validation studies there, since it applies fixed clinical thresholds without a trained model, so the two subsubsections report 27 classification versus 26 supervised studies.

2.4.3.3 Validation method

Validation quantifies a model's generalizability, meaning its behavior on data not used during fitting. In supervised settings, validation estimates the classifier's error on held-out cases; in unsupervised settings, it assesses whether recovered clusters reflect structure that survives beyond the training sample and, where reference labels exist, whether those labels are recovered. Within this corpus, three validation paradigms appear: supervised model validation, unsupervised cluster validation, and statistical or signal-processing studies that fit no predictive model. Table 2.11 reports their distribution across the 49 studies together with their sub-families; the per-study designs are detailed in Appendix A.10.

Paradigm	Sub-family	Studies (n)
Supervised model validation ($n = 26$)	Resampling, no held-out test set	13
	Nested resampling with held-out test set	9
	Single train/test partition	3
	Cross-dataset	1
Unsupervised cluster validation ($n = 13$)	External validity vs. independent labels	6
	Number-of-clusters selection only	6
	No validation reported	1
Statistical / signal-processing ($n = 10$)	No predictive model fitted	10
Total		**49**

Table 2.11: Distribution of the 49 included studies across validation paradigms and sub-families. Within the 26 supervised studies, 10 apply designs that prevent within-subject leakage through LOSO, LOO, or subject-level partitioning, 4 are flagged for known or probable leakage, and 12 describe their partitioning insufficiently to rule it out. Detail per study is provided in Appendix A.10.

A. Conclusion

The three paradigms partition the 49 included studies into 26 supervised, 13 unsupervised, and 10 statistical or signal-processing studies. Within the supervised paradigm, these designs differ in whether a test set is kept separate from the data used for tuning: resampling without a separately held-out test set, where the same folds that tune the model also report its performance, accounts for 13 of 26 studies, nested resampling with a held-out test set, which keeps an outer test set aside for an unbiased estimate while an inner loop tunes, for 9 studies, a single train/test partition, one fixed split without resampling, for 3 studies, and cross-dataset validation, training on one dataset and testing on another, for 1 study. Within the unsupervised paradigm, external validity against independent labels is used by 6 of 13 clustering studies, selection of the number of clusters by 6 studies, and 1 study reports no validation. Ten supervised studies apply designs that prevent within-subject leakage through LOSO, LOO, or subject-level partitioning: subject-level partitioning is the general

rule keeping all of a participant's records on one side of a split, LOSO its per-fold form that holds out one participant at a time, and the LOO designs are here applied at the participant level, so that no participant appears on both sides; four are flagged in the extraction for known or probable leakage [80, 86, 55, 100]; and the remaining twelve describe their partitioning insufficiently to rule out leakage. Methodological comparisons between validation schemes within a single study are rare: two studies from the Sharma group report both K-fold CV and a hold-out split in parallel [82, 83].

2.5 Limitations

This review has methodological limitations that readers should consider when interpreting its conclusions. Several of the limitations listed below could be addressed in future updates of this review.

2.5.1 Search strategy

The Input Signals component of the search query covers modality-level terms such as EEG, ECG, and actigraphy, but omits derived-signal terms such as HRV, RR, oxygen saturation, and skin temperature, all of which appear in the included studies. Studies using these signals were captured only through co-occurrence with other listed terms. A future update of this review could expand the query to include signal-level keywords and assess whether the present corpus missed relevant studies as a result of this omission.

The term "signal processing" was deliberately excluded from the search query to control the volume of retrieved records, as the primary focus was on machine learning and device-based approaches rather than on classical signal processing per se. Studies employing signal-processing methods were not excluded during screening when they were retrieved through other query terms; however, the review does not claim representativeness for pure signal-processing approaches to insomnia assessment.

The search concluded on February 19, 2024. The intervening period has seen rapid development in transformer-based architectures, self-supervised learning, and foundation models for biosignal processing. This review characterizes the methods landscape preceding that shift; a future update would be needed to capture the emerging generation of approaches.

2.5.2 Screening and selection

All 4,184 records were screened by a single reviewer due to resource constraints. The Cochrane Handbook for Systematic Reviews recommends dual independent screening to reduce individual bias [119], and PRISMA 2020 asks authors to report whether this was done [76]. To partially mitigate this limitation, a second reviewer independently screened 5% of the records, totaling 229 references, to estimate inter-rater reliability. The resulting

agreement rate was 98%, with 224 of the 229 records concordant. Five discrepancies were identified and resolved by the primary reviewer. Resolution of discrepancies by the primary reviewer rather than by consensus or a third reviewer is a further deviation from best practice.

2.5.3 Data extraction

All 49 included studies were extracted by a single person with no second-extractor verification on a subsample. The Cochrane Handbook recommends dual independent extraction to reduce transcription errors and interpretation bias [119], and PRISMA 2020 asks authors to report whether this was done [76]. A future iteration of this review could include independent extraction of a random subsample to estimate extraction reliability.

The single extractor made judgment calls documented throughout the chapter, including flagging probable data leakage. These decisions are traceable in the data extraction records but represent one person's interpretation of the source material.

2.5.4 Scope and coverage

The definition of "digital tool" used in inclusion criterion B encompasses methods ranging from analysis of variance on PSG parameters to deep neural networks on raw EEG signals. The resulting methodological heterogeneity, combined with four distinct output types across the 49 included studies, prevented a comparative performance synthesis across methods. These output types are not commensurable, since a binary label, a multi-class assignment, a cluster membership, or a statistical measure cannot be placed on a common performance scale, so no single metric compares them across studies.

2.6 Conclusion

2.6.1 Synthesis of the evidence landscape

The 49 included studies produce four output types: binary classification in 17, multi-class classification in 10, cluster-based phenotyping in 13, and statistical or signal-processing measures in 9. This structural heterogeneity, together with divergent insomnia definitions and non-standardized evaluation protocols, prevented any comparative performance synthesis. Validation weaknesses compound the problem: only one study validates across datasets [68], four are flagged for data leakage, and 12 provide insufficient partitioning detail to rule out within-subject contamination.

2.6.2 Feature design

Traditional sleep parameters align with the parameters used in clinical sleep assessment and are immediately interpretable, but may not capture all features relevant to characterizing insomnia. Mathematical transformations such as entropy, fractal dimensions, and wavelet decomposition capture temporal dynamics that traditional parameters do not describe, but produce features with no direct clinical equivalent.

Several clinically relevant features are underrepresented. No study computes night-to-night variability of HR, total sleep time, or wake after sleep onset, a gap partly explained by the 29 cross-sectional designs of at most 48 hours that preclude across-night analyses. Physical activity beyond circadian actigraphy indices appears in only one study [93], and the apnea-hypopnea index is used as a feature in only two [112, 46]. End-to-end deep learning is adopted by 3 of 49 studies, all on EEG [88, 68, 80], only one study produces probabilistic outputs [112], and no study reports model size or inference time.

2.6.3 Independence of the evidence base

Three research groups account for 11 of the 49 studies: Sharma contributes six, all on the CAP Sleep Database with wavelet-based EEG features [81, 82, 55, 83, 84, 85]; Kusmakar and Angelova co-author three actigraphy studies on different datasets [99, 59, 116]; and Wasiewicz contributes two clustering studies on the same cohort [67, 111]. Eight of 17 public-only studies use the CAP database, six from the Sharma group alone. The frequency of a method in this corpus therefore reflects the interest of individual groups for this topic more than the breadth of its adoption. Similarly, the convergence of five clustering studies on three-cluster solutions [110, 106, 89, 113, 98] does not constitute evidence for a three-subtype model of insomnia, as the five studies define their clusters from entirely different feature spaces.

2.6.4 Directions for future research

Four priorities emerge from these gaps. First, multi-night protocols capturing both sleep and wake periods are needed to satisfy the DSM-5 frequency criterion of at least three symptomatic nights per week and to enable the across-night variability features identified above as underrepresented; the four studies extending beyond two weeks [116, 107, 108, 60] demonstrate feasibility. Second, a shared benchmark dataset with multi-modal recordings and standardized insomnia labels would enable the cross-study comparison that the current corpus cannot support. Third, reporting model size, inference time, and memory requirements should become standard practice for assessing deployment on wearable and nearable devices. Fourth, subjective and objective data should be integrated longitudinally, as one study pairing physiological monitoring with Ecological Momentary Assessment (EMA) over 14 days has demonstrated [107].

2.6.5 Toward phenotyping with multi-night home monitoring

Whether insomnia comprises distinct phenotypes remains unsettled in this corpus: 13 of the 49 studies attempt cluster-based phenotyping, yet the five studies that converge on three-cluster solutions derive them from entirely different feature spaces, so no phenotype structure is established. Resolving this requires evidence the corpus lacks. No study computes the night-to-night variability of HR, total sleep time, or wake after sleep onset, and the 29 cross-sectional designs of at most 48 hours preclude across-night analysis. Home nearables and wearables record many consecutive nights at low cost, which satisfies the DSM-5 frequency criterion of at least three symptomatic nights per week and exposes the night-to-night variability that a single laboratory night cannot show [2]. Several measures could separate phenotypes from such recordings: the variability of the sleep-continuity parameters total sleep time, wake after sleep onset, and sleep efficiency, which may itself be a clinical marker [120]; cardiac and autonomic activity such as HR and HRV as indices of the physiological hyperarousal that characterizes the disorder; objective sleep duration, which distinguishes the objective short-sleep phenotype most strongly tied to cardiometabolic risk [16]; markers of sleep fragmentation; and the apnea-hypopnea index to flag comorbid sleep apnea. Pairing these objective signals with repeated subjective report over time, as the one study combining physiological monitoring with EMA over 14 days demonstrates [107], would test whether phenotypes defined by variability or by sleep duration correspond to insomnia severity and to the risk of harm.

2.7 Chapter Summary

- This review screened 4,184 unique records from four bibliographic platforms and included 49 studies using digital tools for insomnia assessment, with a search concluding in February 2024.
- The 49 studies span four output types: binary classification in 17 studies, multi-class or multi-task classification in 10, cluster-based phenotyping in 13, and statistical or signal-processing analysis in 9.
- 65% of studies use proprietary datasets; the remaining 35% draw from 10 public repositories, with no consensus benchmark.
- 59% of studies rely on a single cross-sectional session of at most 48 hours, precluding assessment of night-to-night variability.
- Subjective data is absent in 78% of public-dataset studies and 41% of proprietary-dataset studies.
- Feature engineering dominates: only 3 of 49 studies use end-to-end deep learning, all on EEG.
- SVM is the most common classifier for binary insomnia classification, appearing in 8 of 10 supervised machine learning studies; k-means is the most common clustering algorithm, appearing in 6 of 13 clustering studies.
- 39 of 49 studies feed features directly to the classifier with no explicit dimensionality reduction.
- Heterogeneous output types, insomnia definitions, and datasets prevent cross-study performance comparison.
- 11 of 49 studies originate from three research groups, and 8 of 17 public-only studies use the same database, meaning the apparent study count overstates independent evidence.
- Extended multi-night monitoring and integration of subjective data as well as objective data are identified as priorities for future research.

Chapter 3

Evaluation of the Sleep Analyzer's sleep stage identification algorithm

3.1 Introduction

Actigraphy is the objective method most commonly used for the longitudinal, home-based monitoring of sleep in insomnia, but it estimates sleep from movement, does not resolve sleep architecture, and tends to overestimate sleep by scoring quiet wakefulness as sleep [37, 38]. The Withings Sleep Analyzer, a contactless under-mattress nearable, offers an alternative that requires no wearable device. Its apnea detection has been validated [49], but its sleep-staging algorithm had not been evaluated against PSG in real-world home conditions. This chapter presents a prospective evaluation of that algorithm in 117 healthy adults who underwent simultaneous home PSG under free-living conditions, assessing agreement with the PSG reference for both sleep-wake distinction, the more important of the two for insomnia assessment, and sleep-stage identification, on the whole cohort and across demographic, clinical, and environmental subgroups. The discussion then situates these results against published consumer-device validations, examines how the recording boundaries and the choice of reference standard affect the reported metrics, and considers the implications for home-based sleep monitoring.

3.2 Paper

Original Paper

Evaluation of a Contactless Sleep Monitoring Device for Sleep Stage Detection at Home in a Healthy Population: Prospective Study in Free-Living Conditions

Marie-Ange Stefanos[1,2,3], MSc; Guillaume de Laboulaye[3], MSc; David Campo[3], PhD; Martin de Gourcuff[3], MSc; Pierre Escourrou[3,4], MD, PhD; Boris Matrot[1], PhD; Anna Sigridur Islind[2], PhD; Pierre Alexis Geoffroy[1,5,6], MD, PhD

[1]Inserm, F-75019, NeuroDiderot, SleepCmd, Université Paris Cité, Paris, France
[2]Department of Computer Science, Reykjavík University, Reykjavik, Iceland
[3]Withings, Issy-les-Moulineaux, France
[4]Centre Interdisciplinaire du Sommeil, Paris, France
[5]Département de Psychiatrie et D'addictologie, Hopital Bichat - Claude Bernard, F-75018, AP-HP, GHU Paris Nord, DMU Neurosciences, Paris, France
[6]GHU Paris - Psychiatry and Neurosciences, Centre ChronoS, Paris, France

Corresponding Author:

Marie-Ange Stefanos, MSc
Withings
2 rue Maurice Hartmann
Issy-les-Moulineaux 92130
France
Phone: 33 141460460
Email: marieangestefanos@gmail.com

Abstract

Background: Sleep is essential for overall health and well-being, but assessing sleep architecture is often costly and time-consuming, relying primarily on polysomnography (PSG). While wrist-worn wearables offer alternatives, they face limitations regarding user compliance, such as battery charging and physical discomfort. Nearable devices address these burdens, but they regularly lack rigorous validation, especially in real-world settings.

Objective: This study evaluates the accuracy and reliability of the Withings Sleep Analyzer (WSA), a contactless sleep monitoring device, compared to PSG in a home setting using a large and diverse cohort of healthy individuals.

Methods: A total of 117 healthy volunteers (69 women; mean 39.9, SD 11.4 years), prospectively recruited from the general population, underwent home-based PSG and simultaneous WSA recording. The study was conducted under free-living conditions, without constraints on substance intake, prebedtime activity, or forced sleep schedules. The main outcomes were the device's performance in sleep-wake distinction and sleep stage identification using accuracy, kappa, sensitivity, specificity, and the mean absolute error of sleep measures on the entire population and demographic, clinical, and environmental subgroups.

Results: WSA demonstrates high sensitivity (93%, 95% CI 92%-94%) for sleep detection and moderate sensitivity (73%, 95% CI 69%-77%) for wakefulness, achieving an overall accuracy of 87% (95% CI 86%-87%) for sleep-wake distinction. The device showed consistent performance across various demographic subgroups, including different age, BMI, mattress, and sleep arrangements (with or without bed partner) categories. Challenges were noted in accurately classifying specific sleep stages, particularly in distinguishing between light and deep sleep, with a mean accuracy of 63% (95% CI 62%-65%) and a Cohen $\varkappa$ of 0.49 (95% CI 0.47-0.51). The WSA tended to overestimate total sleep time (20 min, 95% CI 10 min to 31 min) and light sleep (1 h 21 min, 95% CI 1 h 8 min to 1 h 36 min) while underestimating rapid eye movement (–15 min, 95% CI –23 min to –8 min) and deep sleep (–46 min, 95% CI –59 min to –34 min) durations. Disagreements between expert reviewers were mirrored in part by the WSA's misclassifications. Participants reported altered perceived sleep quality during the night with the PSG, suggesting discomfort during sleep.

Conclusions: Being contactless and placed under the mattress, the WSA offers a promising approach to long-term sleep monitoring in natural home environments. It shows competitive performance in sleep-wake and sleep stage identification

compared to other consumer devices. Progress in wearable and nearable devices is necessary to enhance their accuracy to better support the monitoring of populations with impaired sleep, although limited by an imperfect gold standard. This work also emphasizes the importance of using large, diverse, and challenging datasets, as well as the need for a standardized methodology for accurate sleep stage classification.

JMIR Hum Factors 2026;13:e77033; doi: 10.2196/77033

Keywords: health; informatics; nearable; wearable; sleep stage; home polysomnography; performance assessment; healthy population; sleep monitoring device

Introduction

Sleep is an essential component of human health and plays an important role in daily functioning and overall well-being [1]. Poor sleep quality can significantly impact multiple aspects of life, including physical health, social interactions, and emotional regulation [2,3]. Moreover, poor sleep quality can diminish alertness and cognitive function, increasing the risk of accidents. Furthermore, inadequate sleep has been linked to numerous health issues, such as metabolic disorders, cardiovascular disease, neurocognitive impairments, and mental health challenges [4]. In severe cases, poor sleep quality can lead to critical outcomes, including an increased risk of suicide [5]. Despite the recognized importance of sleep, sleep-related difficulties remain common in the general population [6]. Approximately 30% of adults experience at least one symptom of nocturnal insomnia on a chronic basis [6], over 15% report sleepiness that affects daily activities [7], and more than 10% experience obstructive sleep apnea [8].

Polysomnography (PSG), considered the gold standard for identifying the micro and macro sleep architecture, records a variety of signals (electroencephalogram [EEG], electrooculogram [EOG], electromyogram [EMG], electrocardiogram [ECG], pulse oximetry, air flow, and breathing effort [9]). As the hypnogram helps physicians identify abnormal sleep patterns, quantify sleep fragmentation, and analyze sleep stage distribution, it supports the diagnosis of various sleep disorders, such as sleep-related breathing disorders, narcolepsy, hypersomnia, and disturbing sleep-related behaviors [10].

PSG has limitations. It often involves extensive setup time and significant costs, making it less feasible for routine monitoring [11]. It is scored manually, which is a labor-intensive process that can lead to long waiting times and has been shown to be prone to human error [12]. Additionally, since it is typically conducted over just a few nights, it fails to capture the full picture of an individual's circadian rhythm and night-to-night variability [13,14]. In most cases, PSG is conducted in a sleep lab with numerous sensors that can disrupt natural sleep patterns [15]. For all these reasons, PSG results in data that may not accurately represent a person's typical sleep in their own home [16].

Ambulatory PSG attempts to address these limitations. Lab studies support its accuracy [17-19]. However, in an uncontrolled home setting, sleep metrics may differ; for example, total sleep time (TST) can increase, potentially due to the absence of lab-induced constraints on waking time [20].

The need for more accessible and cost-effective solutions for multinight sleep monitoring remains, driven by the desire to enhance our understanding of sleep quality and facilitate the identification of potential sleep disorders [14,21]. Home-based sleep monitoring devices can enable continuous tracking in the comfort of one's own home, while providing data that reflect real-life conditions more accurately than laboratory-based examinations [22]. This can provide a more realistic picture of an individual's sleep patterns and helps to address the limitations of traditional sleep studies [23].

To address the challenges of cost and complexity in multinight assessments, wearable sleep devices are commonly used to detect sleep and wake states. These devices increasingly combine actigraphy, which tracks body movements [16], with photoplethysmography (PPG) sensors to increase the accuracy of sleep measures estimation [23]. These devices still show two limitations. First, they struggle to accurately differentiate between wakefulness and sleep when individuals remain still while awake [24]. Second, wearable devices require users to maintain proper charging and wearing, leading to frequent data loss and inconsistent performance [25]. In contrast, nearables, which are not in direct contact with the individual, eliminate the need for charging or wearing a device. Nearables for measuring sleep are typically positioned under the mattress or at the bedside and require minimal intervention after setup.

Withings Sleep Analyzer (WSA) is a nearable device designed to provide a noninvasive alternative for monitoring sleep stages over consecutive nights. Despite the proliferation of wearable and nearable sleep-tracking devices, there is a notable lack of rigorous evaluation of the algorithms, leading to concerns about data accuracy and reliability [26,27]. To date, the WSA performance to detect apnea has been evaluated in laboratory conditions [28-32], for which it is European Conformity (Conformité Européenne) marked and Food and Drug Administration–cleared. In addition, it has been assessed for nocturnal sleep periods detection [33] and cardiorespiratory monitoring [34]. However, validation of its sleep stage detection algorithm remains limited. This study, therefore, aims to assess the WSA sleep stage identification algorithm in a real-world home setting in a cohort of 117 healthy adult participants.

There are several important methodological issues addressed in this work. Existing research in sleep-tracking technologies highlights several persistent problems and limitations. Many studies, such as those on the WatchPAT [35] and Somnofy [36] devices, suggest potential utility but often lack comprehensive validation across diverse

populations and real-world settings. Evaluations of consumer-grade devices, like those conducted by Chinoy et al [37] and Kainec et al [38], are frequently limited by small sample sizes and strict exclusion criteria (young individuals with controlled caffeine, alcohol, screen time, and sleep regularity), which limit the generalizability of their findings. These studies typically occur in controlled environments, such as sleep labs, which do not accurately reflect regular user conditions at home.

Moreover, the assessment of device accuracy in these studies is often complicated by two distinct factors. First, the absence of synchronized epoch-by-epoch data in some reports limits performance evaluation to night-scale sleep measures, such as TST, precluding the calculation of detailed classification metrics (eg, accuracy, sensitivity, and specificity). Second, the practice of using lights-off and lights-on times to define the temporal analysis window, which excludes all data outside this interval, impairs the reliability of the computed error for metrics like sleep onset latency. This method indeed assumes observable in-bed time is equivalent to the intention to initiate sleep, which cannot be passively identified in free-living conditions [39].

Recent publications emphasize the importance of assessing these devices under everyday conditions, aligning with our work (eg, [40]). Other research emphasizes the need for standardized evaluation protocols and methodological consistency across studies, reinforcing the approach taken in our research [41-43]. These studies collectively highlight the critical need for standardized and rigorous evaluation methods in sleep monitoring technology. Benchmarking efforts of devices are also being addressed, as recent studies manifest [37,38,44-47].

Our study contributes to the growing literature on sleep tracking by assessing the WSA's performance in a home setting, on a large and diverse sample of healthy individuals, in free-living conditions, that is, without constraints about substance intake, activity before bed, or forced sleeping period, providing insights into the device's accuracy and reliability in real-world conditions. The primary goal is to evaluate WSA's performance using classification metrics and mean absolute error (MAE) on standard sleep metrics [48]. Additionally, we compare performance across subgroups based on demographic and clinical characteristics, such as sex, age, BMI, bed partner presence, mattress type, but also Pittsburgh Sleep Quality Index (PSQI) score [49], mattress, and thickness.

By addressing these gaps, our study enhances the understanding of WSA's capabilities and contributes to the development of more reliable and user-friendly sleep-tracking devices.

Methods

Ethical Considerations

Data were collected during a clinical study approved by the French Institutional Review Board (approval number: 2018-A03129-46) and conducted by Withings from 2020 to 2022. The inclusion process for the study was conducted prospectively by a third party, a general-purpose survey company owning listings of self-declared individuals from the general population. Participants were contacted based on the inclusion and exclusion criteria. In particular, the contacted persons had to have no preexisting medical conditions that would necessitate a PSG examination, and inclusion criteria aimed to ensure a balanced distribution of BMI within each age group. The cohort consisted of a consecutive series of 194 volunteers who provided informed consent. To guarantee data privacy, a unique, randomized anonymization code was assigned to each participant (used for both WSA and PSG data correspondence). Withings only had access to these pseudonymized datasets and never had access to direct identifiers. No compensation from Withings was provided to the participants.

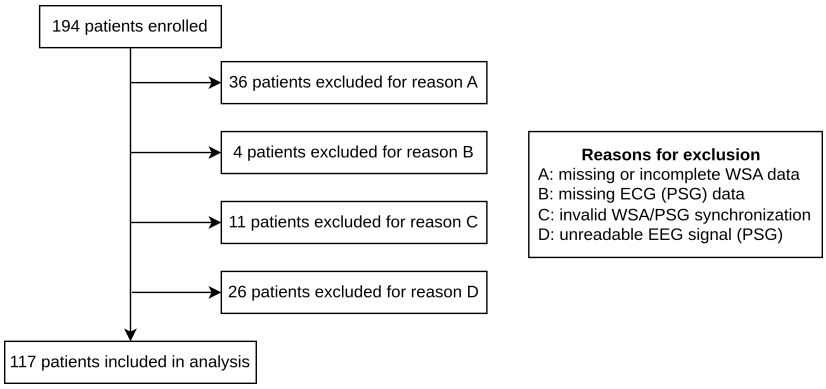
Dataset

Participants

A total of 194 participants slept for the first night at home with WSA and the CID-LXe portable PSG (CIDELEC), both installed by a technician the evening before the recording. The night after that, the participants slept at home with the WSA, without the PSG. Each participant was asked to answer once to the PSQI questionnaire upon enrollment. After each night, participants were asked to complete a survey addressing the following: their sleep quality for that night, a comparison of that night's sleep quality to their usual sleep, whether they took sleeping pills, and any sleep disturbances related to the sensors.

As illustrated in the flowchart (Figure 1), several sources of data loss have led to the inclusion of 117 participants in this analysis. This trial was notably affected by data collection issues due to a dedicated firmware meant to collect high-frequency sensor data (reason for exclusion A). Unlike the commercial version which transmits aggregated data, the devices in this study were programmed to upload raw, high-frequency sensor signals, which introduced connectivity issues detailed in the Limitations section. We performed Little's missing completely at random (MCAR) test, which did not reject the null hypothesis of MCAR data due to reasons A, C, and D ($P>.05$). While a statistically significant result was observed for reason B ($P=.04$), this subgroup comprised only 4 participants and was therefore unlikely to represent a source of systematic bias (see Section A of Multimedia Appendix 1 for detailed results).

Figure 1. Study inclusions flowchart. ECG: electrocardiogram; EEG: electroencephalogram; PSG: polysomnography; WSA: Withings Sleep Analyzer.



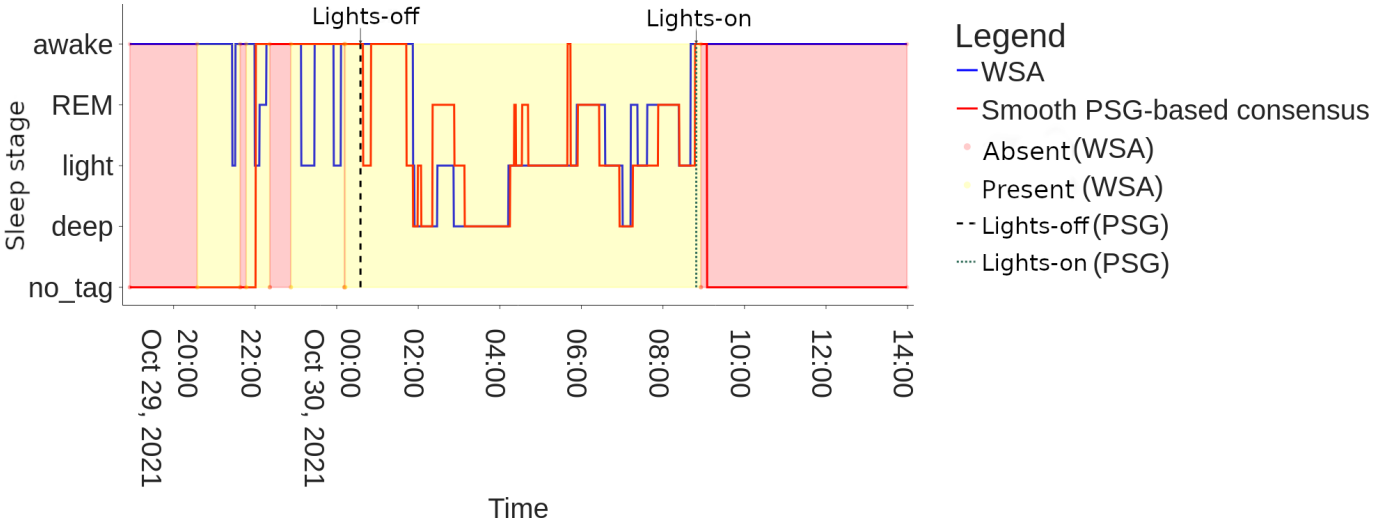
Procedure and Study Design

WSA sampled the raw data from the pressure sensors at 250 Hz and preprocessed data from the microphone at 8 Hz. The reference device was CID-LXe PSG (CIDELEC) that recorded the following signals: 3 EEG, 1 ECG, 2 EOG, 1 submental EMG, and ambient light. Based on experience from past investigations, the digital oximeter and the nasal cannula are among the most sleep-disruptive sensors during PSG, even in a home setting. To minimize this sleep disruption and obtain the most natural possible sleep, PSG was therefore done at home with only EEG, EMG, ECG, and ambient light sensors. As sleep stage annotations from expert reviewers are based on EEG and light sensors only [50], the gold standard reference quality is not affected by the removal of the oximeter and the nasal cannula.

Data Processing

WSA is designed to identify sleep stages over 6-minute windows. As a result, this analysis focuses on evaluating WSA's performance in detecting stages longer than 6 minutes. It should be kept in mind that sleep stages shorter than this are ignored by the device in this study. In addition, WSA does not distinguish stages N1 and N2, so for the aim of comparison, the consensus sleep stages N1 and N2 obtained from the PSG were combined into the "light" category. Rapid eye movement (REM) and awake stages of the WSA and the consensus are directly comparable, and N3 is compared with the stage "deep sleep" from the WSA. The resulting stages are illustrated in Figure 2.

Figure 2. Example of hypnogram (participant 20211029_1) showing WSA sleep stages and the reference (PSG consensus). PSG: polysomnography; REM: rapid eye movement (sleep stage); WSA: Withings Sleep Analyzer.



Two independent certified sleep technologists annotated the PSG signals in 30-second epochs, classifying sleep stages as N1, N2, N3, REM, and awake, following the American Academy of Sleep Medicine (AASM) scoring manual [48]. A third professional reviewed the 2 initial readings and established a consensus for reference data. Sleep stage comparisons were performed over windows of 30 seconds. To evaluate WSA's performance in detecting sleep stages longer than 6 minutes, consensus data were smoothed by selecting for each 6-minute window the stage with the highest occurrence of 30-second epochs. For more details about the procedure, please refer to Section A of Multimedia Appendix 1.

In addition to standard sleep quality measures calculated over the night, such as TST and the duration of each stage, evaluating stage classification required synchronizing

epoch-by-epoch data from both devices over time. Synchronization of the clocks of the devices was obtained by maximizing the correlation of the breathing signals of the two devices [51]. To verify the robustness of our breathing-based synchronization, we assessed the impact of temporal shifts (up to ±90 seconds) on classification performance [39,52]. This sensitivity analysis revealed negligible variations, with the SD for sensitivity and specificity across all sleep stages remaining below 0.005. Details of the procedure can be found in Section A of Multimedia Appendix 1.

An example hypnogram is presented in Figure 2, showing the smoothed, PSG-based consensus in red and WSA sleep stages in blue. WSA indicates bed absence/presence, marked respectively by a red and yellow background, information which is absent from the PSG data. The consensus annotations use instead the events lights-on/off detected by the ambient light sensor of the PSG as proxies for bed-in and bed-out events. For further information on challenges arising from using lights-off/on timestamps as proxies for bed presence, please refer to Section A of Multimedia Appendix 1.

Performance Evaluation Metrics and Tools

Standard sleep quality measures used for this analysis and based on the AASM scoring manual [48] are the following: total sleep time, time spent in each sleep stage (TIREM, TILight, TIDeep, and TIWake), proportion of the time spent in each sleep stage compared to total sleep time (PIREM, PILight, PIDeep, and PIWake), and number of episodes in each sleep stage (NEREM, NELight, NEDeep, and NEWake) [50]. Each of these measures has been computed by each device (PSG and WSA) and then compared using the MAE and a Bland-Altman analysis, described in Section B of Multimedia Appendix 1.

Epoch-by-epoch agreement is assessed using standard classification metrics. Subgroup analyses were conducted to determine whether WSA performance is affected by demographic, clinical, and environmental factors. Details regarding the evaluation metrics and subgroup analyses are available in Sections B and C of Multimedia Appendix 1.

To compare the median MAE between different groups (categories of a given metadata as listed in Table S3 in Multimedia Appendix 1), a Kruskal-Wallis test has been performed on the MAE obtained on all the metadata described in Table S3 in Multimedia Appendix 1 for each of the following sleep quality measures (TST, TIREM, TILight, TIDeep, PIREM, PILight, and PIDeep) and for the number of episodes in each sleep stage.

To compare the MAE variance between different groups, a Brown-Forsythe test has been performed on the same data for the same sleep quality measures. The Holm-Bonferroni correction is used after both statistical tests to address the issue of multiple comparisons.

Results

Patient Characteristics

Demographic and clinical data are summarized in Table 1 for the population included in the analyses.

Table 1. Demographic characteristics and clinical data for the entire population (N=117).

Characteristics	Values
Demographics	
Age (years)	
Mean (SD)	39.9 (11.4)
Min-max	19-69
Height (cm)	
Mean (SD)	171.6 (8.3)
Min-max	153-194
Weight (kg)	
Mean (SD)	70.4 (12.0)
Min-max	49-103
BMI (kg/m^2)	
Mean (SD)	23.9 (3.3)
Min-max	17.9-30.9
Sex, n (%)	
Male	48 (41)
Female	69 (59)
Skin type[a], n (%)	
Skin type 1	85 (75)
Skin type 2	21 (18)
Skin type 3	8 (7)

Characteristics	Values
Known pathologies, n (%)	
Asthma	5 (4)
Arterial hypertension	3 (<3)
Bipolar disorder	1 (<1)
Ulcerative colitis	1 (<1)
Ankylosing spondylitis	1 (<1)
Pathologies, n (%)	
Cardiovascular[b]	4 (4)
Neurological[c]	0 (0)
Digestive[d]	0 (0)
Metabolic[e]	2 (2)
Psychiatric[f]	0 (0)
Other	6 (5)
PSQI[g] questionnaire [49]	
PSQI score	
Mean (SD)	6.0 (2.9)
Min-max	0-15
Component 1: subjective sleep quality	
Mean (SD)	1.2 (0.7)
Min-max	0-3
Component 2: sleep latency	
Mean (SD)	1.3 (0.9)
Min-max	0-3
Component 3: sleep duration	
Mean (SD)	0.8 (0.8)
Min-max	0-2
Component 4: habitual sleep efficiency	
Mean (SD)	0.4 (0.7)
Min-max	0-3
Component 5: sleep disturbances	
Mean (SD)	1.3 (0.5)
Min-max	0-3
Component 6: use of sleeping medications	
Mean (SD)	0.2 (0.6)
Min-max	0-3
Component 7: daytime dysfunction	
Mean (SD)	0.9 (0.8)
Min-max	0-3
Medications, n (%)	
Antihypertensive	3 (3)
Antidepressant	2 (2)
Hypnotics	1 (1)
Corticoid	1 (1)
Antiepileptic	1 (1)
Antipsychotic	1 (1)
Mattress type, n (%)	
Foam	41 (35)
Spring	29 (25)
Latex	24 (21)

Characteristics	Values
Memory foam	12 (10)
Other	8 (7)
Unknown	3 (3)
Bed partner, n (%)	
Yes	64 (55)
No	52 (45)
Sensors disturbance (due to PSG[h] device), n (%)	
Yes	46 (39)
No	71 (61)

[a]Skin type is based on the reduced Fitzpatrick scale.
[b]Arterial hypertension, atrial fibrillation, or myocardial infarction.
[c]Parkinson, stroke, adenoma, manic-depression, or epilepsy.
[d]Hemolytic anemia due to glutathione reductase deficiency, ulcer, sleeve, or diverticula.
[e]Diabetes, hypothyroidism, dyslipidemia, or obesity/overweight.
[f]Depression, insomnia, or anxiety.
[g]PSQI: Pittsburgh Sleep Quality Index.
[h]PSG: polysomnography.

A total of 117, mostly healthy, participants were included in the analysis. The cohort comprised 59% (n=69) women and 41% (n=48) men, with a mean age of 39.9 (SD 11.4) years, ranging from 19 to 69 years. The average BMI was 23.9 (SD 3.3) kg/m².

The average sleep quality in this population is considered mildly impaired, with a mean PSQI score of 6 out of 21. The SD of 2.9 highlights significant variability.

The most common mattress types were foam (n=41, 35%), spring (n=29, 25%), and latex (n=24, 21%). Around 55% (n=64) of participants reported having a bed partner. Sensor disturbances were reported by 46 (39%) participants due to the PSG device, while no disturbances were reported regarding the WSA nearable.

Sleep Stages Classification

Figure 3 presents the confusion matrix containing the epochs across all nights, and Table 2 gathers sleep stages classification results on the whole dataset.

Figure 3. Sleep stage classification confusion matrix (N=117). (Left) Raw confusion matrix displaying total epoch counts for WSA sleep stage detection versus PSG ground truth. (Right) Row-normalized confusion matrix (expressed in percentages) illustrating the classification sensitivity and misclassification rates for each true sleep stage. PSG: polysomnography; REM: rapid eye movement (sleep stage); WSA: Withings Sleep Analyzer.

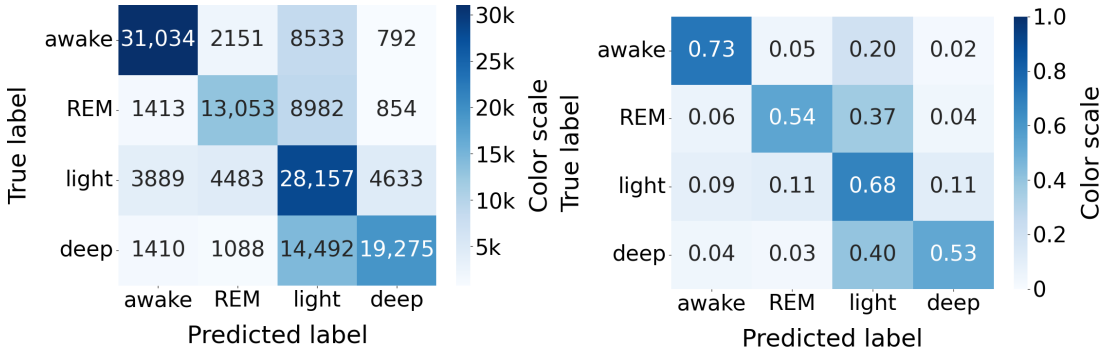


Table 2. Summary of sleep stage classification results over nights (N=117) (computed as the averages of all nights).

Mean of results per night	Sleep stages								Sleep-wake			
	Kappa	Accuracy	Se[a] REM[b]	Se light	Se deep	Sp[c] REM	Sp light	Sp deep	Kappa	Accuracy	Se sleep	Sp sleep
Mean	0.49	0.63	0.53	0.68	0.56	0.93	0.69	0.94	0.65	0.87	0.93	0.73
95% CI	0.47-0.51	0.62-0.65	0.48-0.57	0.65-0.71	0.51-0.60	0.92-0.94	0.67-0.71	0.93-0.95	0.62-0.69	0.86-0.89	0.92-0.94	0.69-0.77
SD	0.12	0.09	0.24	0.16	0.25	0.05	0.11	0.06	0.19	0.07	0.06	0.22

[a]Se: sensitivity.
[b]REM: rapid eye movement (sleep stage).

[c]Sp: specificity.

Sensitivity and Specificity

The WSA sensitivity for detecting sleep was 93% (95% CI 92%-94%; N=117), meaning it correctly identifies when a person is asleep 93% of the time. The mean specificity was 73% (95% CI 69%-77%), indicating it accurately detects when a person is awake 73% of the time.

The performance in accurately classifying each individual sleep stage against the other three is more variable, with sensitivities of 53% (95% CI 48%-57%), 68% (95% CI 65%-71%), 56% (95% CI 51%-60%), and 73% (95% CI 69%-77%) for REM, light sleep, deep sleep, and wake stage, respectively. The corresponding specificities were 93% (95% CI 92%-94%), 69% (95% CI 67%-71%), 94% (95% CI 93%-95%), and 93% (95% CI 92%-94%), respectively.

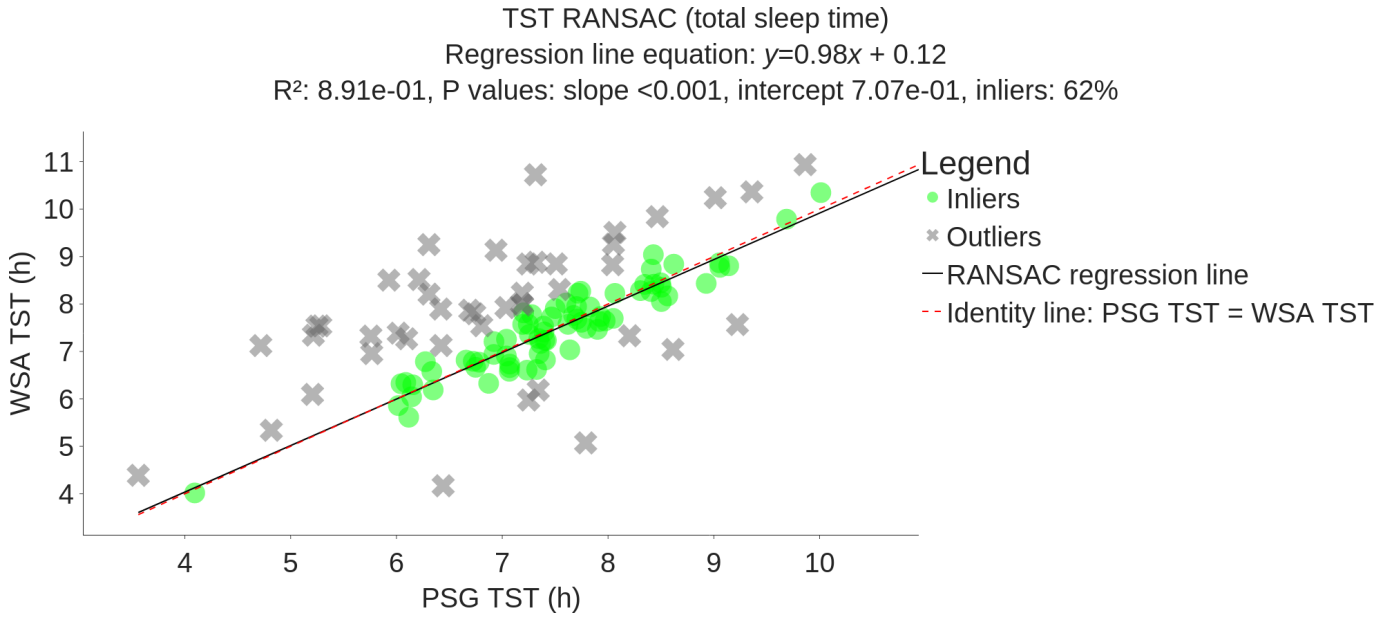
Cohen κ and Accuracy

Please note that, contrary to sensitivity and specificity, Cohen κ and accuracy depend on the relative sizes (prevalences) of each class. In this study's dataset, the WSA mean Cohen κ was 0.65 (95% CI 62%-69%) for distinguishing between sleep and wake stages, and 0.49 (95% CI 47%-51%) for differentiating among the 4 stages (N=117). WSA accuracy to distinguish between sleep and wake was 87% (95% CI 86%-89%). The mean accuracy for the 4 stages was 63% (95% CI 62%-65%).

Assessment of Sleep Quantity Measures

Figure 4 shows the TST estimated by the WSA versus the PSG. The slope of the linear regression is 0.98 (P<.001) and the R^2 value is 0.89, with 38% of outliers, indicating a moderate linear relationship with PSG.

Figure 4. RANSAC regression plot between WSA and PSG total sleep time (black line) compared to the identity line (red dashed line) on 117 participants (R^2=0.89, slope P value P < .001, inliers: 62%). PSG: polysomnography; RANSAC: random sample consensus algorithm; TST: total sleep time; WSA: Withings Sleep Analyzer.



WSA deep sleep was not correlated to time spent in deep sleep (slope estimate of 0.05, P=.7, Figure S4 in Multimedia Appendix 1). The WSA estimates of light sleep and REM were moderately correlated with PSG values, with respective slopes of the regression lines of 0.4 (P<.001, Figure S5 in Multimedia Appendix 1) and 0.65 (P<.001, Figure S6 in Multimedia Appendix 1).

Precision of the WSA estimates is measured with the MAE. The MAE values of total sleep time (equal to the MAE of wake duration because the sum of TST and wake duration is the recording time), REM duration, light sleep duration, and deep sleep duration are 44 min, 35 min, 1 h 33 min, and 1 h 4 min, respectively, with average PSG values of 7 h 15 min (3 h 03 min for awake duration), 1 h 43 min, 2 h 57 min, and 2 h 35 min, respectively. Corresponding 95% CIs are reported in Table 3. The average number of episodes in each sleep stage and the average number of stage changes can be found in Table 4. The WSA tends to underestimate the number of episodes for each sleep stage and is less sensitive to transitions between epochs.

Table 3. Summary of Bland-Altman results obtained on sleep-related metrics between WSA[a] and PSG[b] estimations (N=117).

Sleep quality measures	PSG value, mean (SD)	MAE[c] (95% CI)	Bias, WSA – PSG (95% CI)
TST[d]	7 h 15 min (1 h 14 min)	44 min (36 min to 52 min)	20 min (10 min to 31 min)
TIREM[e]	1 h 43 min (45 min)	35 min (30 min to 40 min)	–15 min (–23 min to –8 min)
TILight[f]	2 h 57 min (1 h 12 min)	1 h 32 min (1 h 21 min to 1 h 44 min)	1 h 21 min (1 h 8 min to 1 h 36 min)
TIDeep[g]	2 h 35 min (54 min)	1 h 4 min (55 min to 1 h 13 min)	–46 min (–59 min to –34 min)
TIWake[h]	3 h 3 min (1 h 19 min)	44 min (36 min to 52 min)	–20 min (–31 min to –10 min)
PIREM[i]	17% (8.8%)	5.7% (4.9% to 6.5%)	–2.4% (–3.6% to –1.1%)
PILight[j]	29% (14%)	15% (13% to 17%)	13% (11% to 15%)
PIDeep[k]	25% (11%)	10% (8.8% to 12%)	–7.4% (–9.4% to –5.4%)
PIWake[l]	29% (11%)	7.1% (5.8% to 8.4%)	–3.3% (–5.0% to –1.6%)

[a]WSA: Withings Sleep Analyzer.
[b]PSG: polysomnography.
[c]MAE: mean absolute error.
[d]TST: total sleep time.
[e]TIREM: time spent in rapid eye movement sleep stage.
[f]TILight: time spent in light sleep stage.
[g]TIDeep: time spent in deep sleep stage.
[h]TIWake: time spent in wake stage.
[i]PIREM: proportion of the night in rapid eye movement sleep stage.
[j]PILight: proportion of the night in light sleep stage.
[k]PIDeep: proportion of the night in deep sleep stage.
[l]PIWake: proportion of the night in wake stage.

Table 4. Comparison of the number of sleep stage episodes and stage changes between WSA[a] and PSG[b] (N=117).

Sleep quality measure	Episodes[c] of REM[d]	Episodes[c] of light	Episodes[c] of deep	Episodes[c] of wake	Stage changes
WSA, mean (95% CI)	4.1 (3.8-4.5)	8.9 (8.4-9.3)	4.4 (4.0-4.7)	4.4 (4.1-4.8)	8.9 (8.4-9.3)
PSG, mean (95% CI)	5.0 (4.7-5.3)	13.2 (12.7-13.8)	7.1 (6.7-7.6)	4.9 (4.5-5.4)	13.2 (12.7-13.8)

[a]WSA: Withings Sleep Analyzer.
[b]PSG: polysomnography.
[c]An episode is a sequence of consecutive epochs in the same sleep stage.
[d]REM: rapid eye movement (sleep stage).

Additional scatter plots and regression analyses of these metrics can be found in Section D of Multimedia Appendix 1.

The Bland-Altman plots are presented in Figures 5 and 6. Biases are summarized in Table 3.

Figure 5. Bland-Altman plot comparing WSA and PSG estimation of total sleep time and time in REM on 117 participants. LOA: limits of agreement; PSG: polysomnography; REM: rapid eye movement (sleep stage); TIREM: time spent in rapid eye movement sleep stage; TST: total sleep time; WSA: Withings Sleep Analyzer.

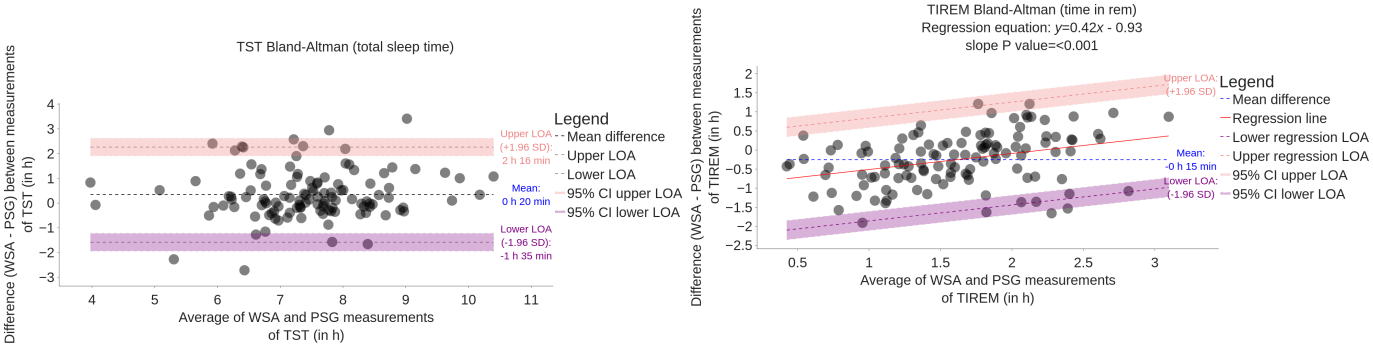
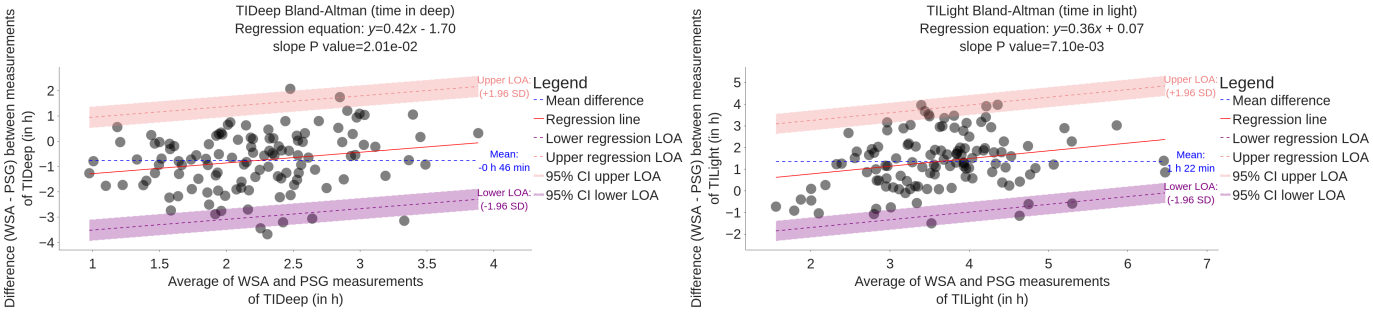


Figure 6. Bland-Altman plot comparing WSA and PSG estimation of time in deep and time in light (N=117). LOA: limits of agreement; PSG: polysomnography; TIDeep: time spent in deep sleep stage; TILight: time spent in light sleep stage; WSA: Withings Sleep Analyzer.



In Figure 5, the Bland-Altman plot for TST does not reveal any proportional bias with the duration of the night. The mean error is 20 minutes. The wide limits of agreement show that there is a substantial amount of variability between PSG and WSA. By contrast, the average error in REM, light, and deep sleep durations depends on the duration of the phase, therefore exhibiting proportional bias. Light sleep is *on average* overestimated for any duration, deep sleep is *on average* underestimated for any duration, and REM is *on average* underestimated for short durations and overestimated for long durations.

The mean error (bias) in light sleep duration is important, with 1 h 21 min. Biases in REM and deep sleep are –15 min and –46 min, respectively. The mean error in TST is therefore 20 min.

Subgroup Analyses

No significant differences were found in the median or variance of the error across most groups for all the quantitative sleep variables of Table 3. This includes comparisons based on gender, BMI categories, age groups, sleeping arrangements, PSQI scores, self-reported sleep quality, mattress thickness, and night awakenings. However, significant differences in median total sleep time and wakefulness proportion were observed within the PSQI sleep duration component subgroup. Median MAE of TST is respectively 21 min and 1 h 11 min for good and bad sleep duration subgroups, and median MAE of wakefulness proportion is respectively 3% and 12% for good and bad sleep duration subgroups (N=117). This finding suggests that distinguishing sleep from wake is a harder task for the WSA when the participant has subjective short sleep duration (<6 h), highlighting an area for further investigation. For a comprehensive overview of the statistical tests and detailed results, please refer to Multimedia Appendix 2.

As an illustration, Figures 7-9 show the boxplots of TST MAE of WSA compared to PSG, stratified by age groups, BMI ranges, and whether the participant was alone in bed. They suggest that the WSA performance did not depend on these factors. Boxplots for TST MAE for the other subgroups can be found in Section D of Multimedia Appendix 1.

Figure 7. Boxplot of TST mean AE (hours) by age group (N=117). AE: absolute error; TST: total sleep time.

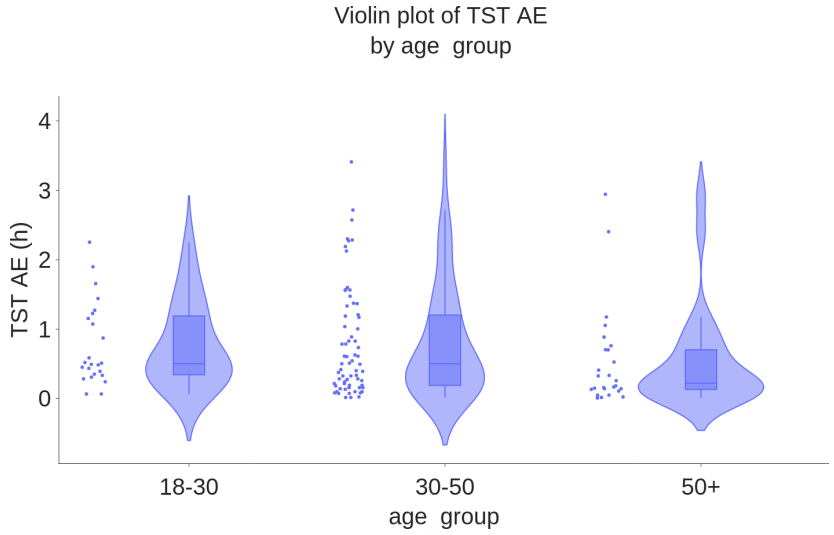


Figure 8. Boxplot of total sleep time mean AE (hours) by BMI group (N=117). AE: absolute error; TST: total sleep time.

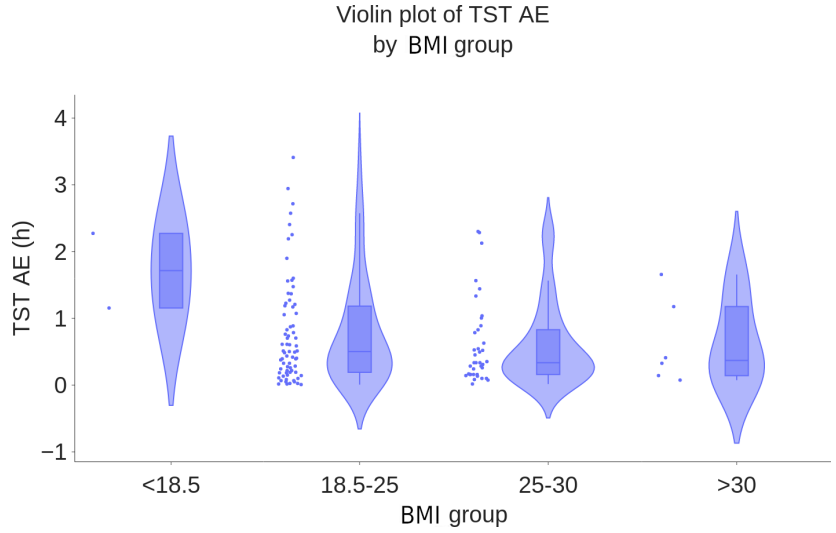


Figure 9. Boxplot of total sleep time mean AE (hours) by *is couple* group (N=117). No: participants who slept alone the night of the recording; yes: participants who had a sleep partner the night of the recording. AE: absolute error; TST: total sleep time.

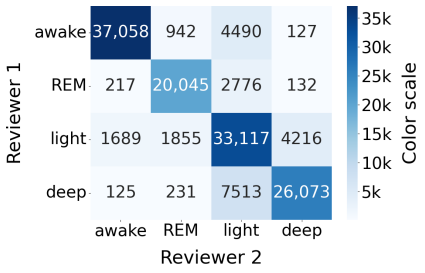


Agreement Between Reviewers

Each 30-second epoch of the PSG data was independently annotated by two reviewers. The mean κ value for sleep stage determination between reviewers before reaching consensus is 0.71 (SD 0.11), closely aligning with literature findings (κ=0.76 in [53]), while it is 0.81 (SD 0.17) for sleep-wake classification.

In the analysis of reviewer agreement on sleep stage classifications, the primary disagreements occur between light and deep stages, as well as between light sleep and awake/REM stages (Figure 10). This indicates that even expert reviewers, considered the gold standard, face challenges in accurately distinguishing these stages due to their inherent similarities and subtle physiological differences. This inherent difficulty in distinguishing light, deep, awake, and REM sleep stages by human experts is reflected in WSA, which was trained on PSG data labeled using a similar protocol. It is likely that the imprecision of the gold standard is passed on to the device.

Figure 10. Sleep stage classification confusion matrix between the first two reviewers' annotations (N=117). REM: rapid eye movement (sleep stage).



Discrepancies in Performance: Epochs With Reviewers' Agreement vs Disagreement

Given the discrepancies encountered in PSG annotation among reviewers, it is pertinent to investigate whether the WSA performed better on epochs where reviewers reached consensus compared to those where disagreements were present. To perform this analysis, the dataset was divided

into two groups of epochs: those where reviewers reached agreement and those where disagreements occurred. Separate performance analyses were then conducted on these two complementary groups.

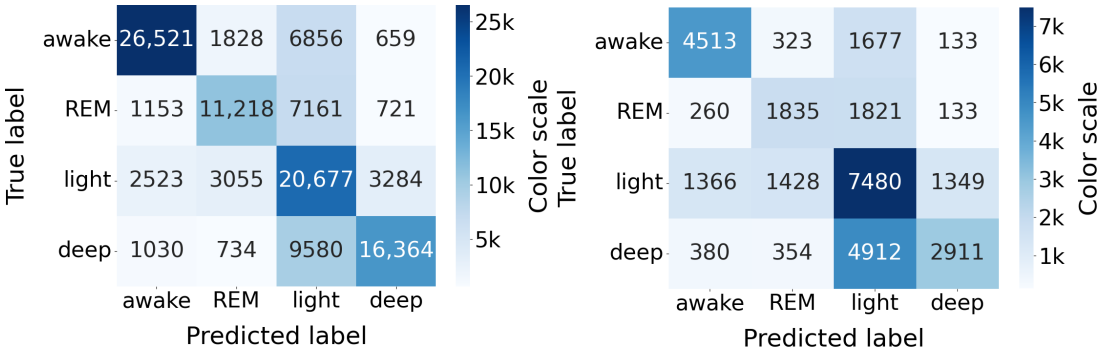
Results are summarized in Table 5. In the "agreement" column (resp. "disagreement" column), only epochs where both reviewers agreed (resp. disagreed) were included, while epochs with disagreement (resp. agreement) were excluded. Figure 11 shows the confusion matrices obtained for both sets.

Table 5. Summary of performance in reviewers' agreement and disagreement subgroups (N=117).

WSA[a] vs PSG[b]	Agreement	Disagreement	*z* score
Sleep stages			
Epochs, n (%)	113,364 (79)	30,875 (21)	—[c]
Kappa, mean (95% CI)	0.52 (0.50-0.55)	0.27 (0.24-0.30)	78.06
Accuracy, mean% (95% CI)	0.66 (0.64-0.68)	0.53 (0.50-0.55)	42.00
Se[d] REM[e], mean (95% CI)	0.55 (0.50-0.59)	0.44 (0.39-0.49)	34.32
Se light, mean (95% CI)	0.7 (0.67-0.72)	0.64 (0.60-0.67)	20.16
Se deep, mean (95% CI)	0.56 (0.51-0.60)	0.37 (0.32-0.43)	68.80
Se wake, mean (95% CI)	0.72 (0.67-0.76)	0.51 (0.44-0.58)	69.85
Sp[f] REM, mean (95% CI)	0.94 (0.93-0.95)	0.92 (0.90-0.93)	12.70
Sp light, mean (95% CI)	0.72 (0.69-0.74)	0.52 (0.48-0.56)	66.63
Sp deep, mean (95% CI)	0.95 (0.94-0.96)	0.92 (0.90-0.94)	20.25
Sp wake, mean (95% CI)	0.94 (0.93-0.95)	0.92 (0.90-0.93)	12.70
Sleep-wake			
Epochs, n (%)	133,447 (93)	10,792 (7)	—
Kappa, mean (95% CI)	0.64 (0.59-0.68)	0.33 (0.26-0.38)	63.72
Accuracy, mean% (95% CI)	0.88 (0.86-0.89)	0.76 (0.71-0.78)	35.77
Se wake, mean (95% CI)	0.74 (0.67-0.76)	0.52 (0.46-0.59)	49.15
Se sleep, mean (95% CI)	0.94 (0.93-0.95)	0.86 (0.83-0.89)	32.20

[a]WSA: Withings Sleep Analyzer.
[b]PSG: polysomnography.
[c]Not applicable.
[d]Se: sensitivity.
[e]REM: rapid eye movement (sleep stage).
[f]Sp: specificity.

Figure 11. Sleep stage classification confusion matrices (N=117) on sets with agreement (left) and disagreement (right); WSA sleep stage detection vs PSG ground truth. PSG: polysomnography; REM: rapid eye movement (sleep stage); WSA: Withings Sleep Analyzer.



We used a *z* test for proportions to test for differences between sensitivity, specificity, accuracy, and kappa. All the results are statistically significant, with WSA performing better on epochs where the two reviewers agreed. The progression is particularly significant for the sensitivity in deep sleep, reaching 0.56, compared to 0.37 when the

reviewers disagreed, as in the latter case, WSA predominantly classified deep sleep stages as light sleep. The specificity for light sleep was also starkly improved from 0.52 to 0.72. As a result, the Cohen $\varkappa$ is almost double when reviewers agreed (0.64 vs 0.33 for S/W classification, and 0.52 vs 0.27 for 4-stage classification).

Sleep Quality With vs Without PSG

Since this analysis does not involve the WSA or the PSG data, but only the participants' answers to the questionnaires, it was conducted on the 186 participants for whom reported sleep quality data were available. Results are given in Table 6.

Table 6. Comparison of reported sleep quality and sensor disturbance with and without PSG[a] (n=186). In both cases, the participant used a WSA[b] and two ScanWatch devices.

Qualitative sleep quality measures	First night with PSG	Second night without PSG
Sleep quality, n (%)		
Very poor	10 (5.4)	3 (1.6)
Poor	65 (34.9)	32 (17.2)
Quite good	75 (40.3)	71 (38.2)
Good	32 (17.2)	58 (31.2)
Very good	4 (2.2)	22 (11.8)
Sleep quality compared to usual sleep quality, n (%)		
Much worse	28 (15.1)	6 (3.2)
Worse	69 (37.1)	32 (17.2)
Same	77 (41.4)	111 (59.7)
Better	10 (5.4)	31 (16.7)
Much better	2 (1.1)	6 (3.2)
Sensor disturbance		
Device that caused sleep disturbances	PSG	WSA
Yes, n (%)	77 (41.4)	0 (0)
No, n (%)	109 (58.6)	0 (0)

[a]PSG: polysomnography.
[b]WSA: Withings Sleep Analyzer.

Despite the removal of the nasal cannula and the blood oximeter from our home PSG test, the sleep test was associated with significant discomfort in 41% (77/186) of the participants during the night of the PSG home test, compared to 0% (0/186) in the subsequent night with the WSA only. Fifty-two percent (97/186) (resp. 20%, 38/186) of the participants reported a degraded sleep quality on the night of the PSG (resp. without the PSG).

Figure 12 illustrates the changes in sleep quality among participants over 2 consecutive nights. Participants generally experienced better sleep without the PSG sensors. Notable improvements were observed in 52% (97/186) of cases, with transitions from fairly good to good (26/186, 14%), and from poor to fairly good (31/186, 17%), good (12/186, 6%), or very good (10/186, 5%). Additionally, some participants reported significant improvements from very poor to very good sleep, and others from poor to fairly good. These improvements may be influenced by several factors, including the removal of PSG sensors, which can be intrusive and uncomfortable, potentially disrupting sleep. Furthermore, the awareness of being recorded might alter participants' sleep behavior, either positively or negatively. The use of unfamiliar devices could also play a role, as participants may initially experience discomfort or anxiety, which diminishes over time. External conditions such as noise and room temperature, which were not controlled in this study, could also significantly impact sleep quality. Therefore, while the data suggest improvements, it is crucial to consider these potential confounding factors when interpreting the results.

Figure 12. Comparison of reported sleep quality with and without PSG (N=186). Green: improved; yellow: identical; red: deteriorated; black: first night with PSG; grey: second night without PSG. In both cases, the participant used a WSA and two ScanWatch devices. PSG: polysomnography; WSA: Withings Sleep Analyzer.

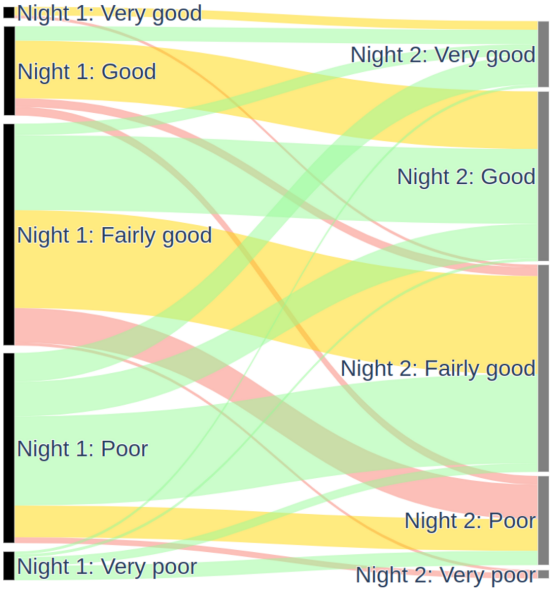


Table 6 highlights that participants slept as well with WSA as they typically do. On the second night, 59.7% (111/186) of participants reported the same sleep quality, 16.7% (31/186) reported better sleep, and 3.2% (6/186) reported much better sleep compared to their usual experience. This suggests that Withings sleep monitoring devices do not negatively impact sleep quality and may even enhance it for some users.

Discussion

Main Findings

The evaluation of the WSA for sleep stage detection in a cohort of 117 participants in a home setting revealed several key insights. The device demonstrated a high sensitivity of 93% for detecting sleep, underscoring its strong capability to accurately identify sleep periods. However, the specificity was moderate at 73%. The duration of the recording before actual sleep onset and after wake-up significantly affects observed performance of wake detection, with shorter periods inflating specificity, accuracy, and Cohen κ. This occurs due to a shortening of quiet wake periods hard to distinguish from sleep (reading or screen time in bed, for instance), leading to fewer false positives while true negatives remain relatively constant. Conversely, overly long (eg, 24-hour) recordings, dominated by easily identifiable diurnal wakefulness, also tend to inflate specificity. In both cases, the performance of the evaluated device is overestimated. The most clinically relevant and challenging scenario for performance evaluation involves analyzing data encompassing the entire in-bed period, including calm wakefulness before sleep onset and after awakening. While AASM guidelines suggest using lights-off and lights-on as proxies for night boundaries, this approach truncates crucial pre– and post–sleep wake data, therefore introducing a bias in TST and classification performance endpoints. To mitigate this, we used all available PSG annotations, irrespective of lights-off and lights-on times, to provide a more comprehensive and unbiased assessment of performance across the full in-bed duration.

The classification of sleep stages (into wake, REM, light, and deep phases) is more challenging. The confusion matrix (Figure 3) shows that 40% of deep sleep phases, 37% of REM sleep, and 20% of wake phases were misclassified as light sleep by the WSA. As a result, the corresponding sensitivities were rather low (0.56, 0.53, and 0.73, respectively). Thirty-two percent of light sleep stages were misclassified, in similar proportions, into REM, deep sleep, and wake. In consequence, (1) the REM and deep sleep total durations were underestimated (by approximately 30% and 15%), while light sleep duration was overestimated by almost 48% (biases in Table 3), and (2) the specificities for wake, REM, and deep sleep are very high (0.93, 0.94, and 0.94, respectively), while the specificity for light sleep is moderate (0.69).

We showed that these misclassifications of WSA mirror the disagreements between PSG reviewers, who mainly disagree on light versus all the other sleep stages (Figure 11). WSA classification was indeed significantly better on epochs where reviewers agreed (79% of the epochs) than on epochs where they disagreed. WSA's sleep stage classification uses an algorithm trained by supervised learning on PSG data annotated by technician experts, so it is expected that WSA performance is limited by the interreader variability.

Consistent with existing validation studies [33,44], we also report accuracy and Cohen κ. We caution that these performance metrics are strongly dependent on the prevalence of the classes (here, wake, sleep, REM, deep, and light), which creates difficulties in comparing results between studies. In particular, for sleep studies, two factors modify the prevalence of the classes significantly. First, there is no standardized definition of the beginning of the night. Second, the choice of the reference device has some importance, causing more or less sleep disturbances.

With these caveats in mind, the WSA achieved an overall accuracy of 87% for sleep-wake distinction, which is promising for a contactless home monitoring device. Despite this, the accuracy for classifying specific sleep stages was moderate, with REM sleep at 53%, light sleep at 68%, and deep sleep at 56%. The overall accuracy for sleep stage classification was 63%, with a Cohen κ of 0.49.

Notably, the WSA's performance was consistent across various demographic subgroups, including different age and BMI categories. Importantly for a device under the mattress, there was no difference between individuals who slept alone and those with a bed partner, nor between different types or thicknesses of mattresses. Other groups analyzed included sex, PSQI scores, self-reported sleep quality, and whether individuals woke up during the night, all of which also showed no significant differences in MAE, except for wake proportion by subjective sleep duration group. This consistency suggests the device may be useful for a wide range of participants.

Comparison With Previous Research

Differences in Studies' Protocols

To contextualize our findings, we used published validation studies of other consumer sleep trackers. We advise interpreting these comparisons with caution, as study protocols vary significantly (eg, in-lab vs at-home settings, population characteristics, and data handling procedures). We compare our results on WSA with influential studies that have contributed to the understanding of consumer sleep-tracking technology. Studies [29], [37], and [33], respectively, evaluated 7, 5, and 6 consumer devices against PSG, mainly wearable devices, revealing significant variability in their accuracy for sleep stage detection. Studies [33,54-60] compared a total of 8 contact-free devices and 1 wearable against PSG. These studies provide a valuable context for evaluating our device's performance, illustrating both the advancements and challenges in the field of consumer sleep technology. All the data used for this comparison can be found in Multimedia Appendix 3.

We compared the performance metrics reported by other studies on their respective datasets to the performance we

obtained on our dataset, without reproducing their results or running their device algorithms on our data. When a result was not explicitly provided in the study report, we calculated it from the confusion matrix if it was available.

Consistent with recent validations [39], our study population encompasses a diverse range of cofactors of sleep quality, presenting a wide range of age and BMI, as was the case in only 2 of the 11 studies used for the comparison [55,59]. Four studies [37,38,44,58] included only young adults, and one study focused on older adults [33].

The conditions of our study are not controlled since participants stayed at home, as in only one other study [54] but on a much smaller scale (n=20). Our dataset is therefore richer regarding important environmental cofactors of sleep quality, including sleeping arrangements (alone or with a partner), mattress types, ambient temperature, lighting and sound, prebedtime activities (eg, no restriction on caffeine or alcohol consumption, on screen time, no forced sleep regularity), and activities during time in bed (eg, no forced turning off of the lights).

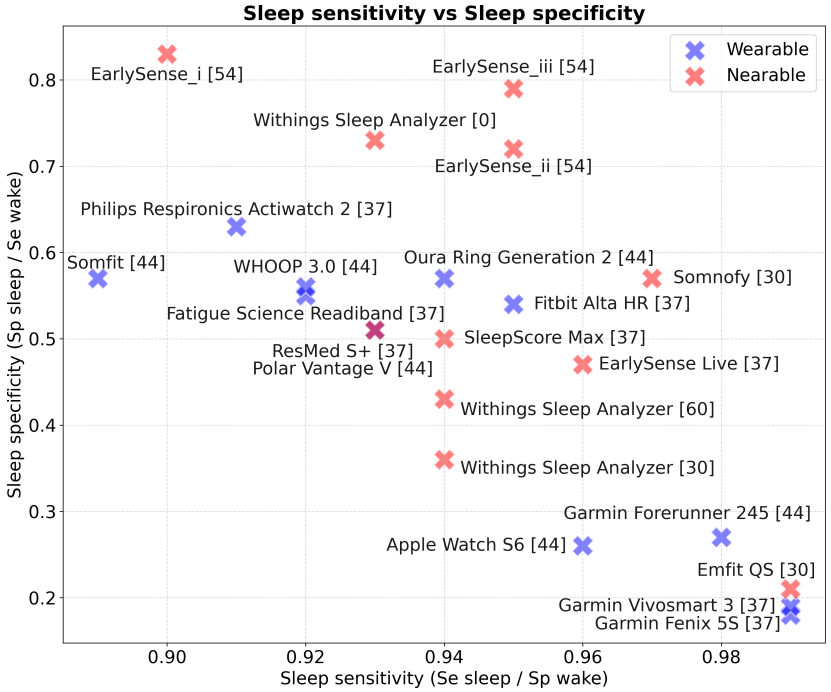
In addition, some studies use strict exclusion criteria also after data collection [38] to ensure data quality and reliability, such as excluding individual nights with a total recording time of less than 7 hours and 50 minutes [37].

Thus, our real-world home setting captures the natural variability in sleep patterns and environmental factors. While this enhances the generalizability of our results, reflecting typical user experiences, it also presents a less favorable scenario for the device, posing additional challenges for accurate sleep stage detection and analysis (for instance, activities in bed like reading or screen time can be confounded by the device as sleep). Consequently, our findings provide a more realistic evaluation of the device's performance in everyday conditions.

Sleep-Wake Classification

The choice of a sleep monitoring device requires being informed about the trade-off between sleep and wake identification. This is illustrated in Figure 13. Devices with high sleep sensitivity (≥0.98) exhibit low specificity (<0.30), as observed with the Garmin Forerunner 245, Emfit QS [33], Garmin Vivosmart 3, and Garmin Fenix S5. Conversely, devices achieving higher specificity (≥0.70), such as the EarlySense [54] and WSA, tend to show slightly lower sensitivity (0.90-0.95).

Figure 13. Overview of sleep-wake reported sensitivities from varying validation studies. Indication [0] denotes this study. Note: Direct comparison is limited by differences in study protocols and populations [33,37,44,54,60]. The exact data points corresponding to the figure are fully detailed in Multimedia Appendix 3.



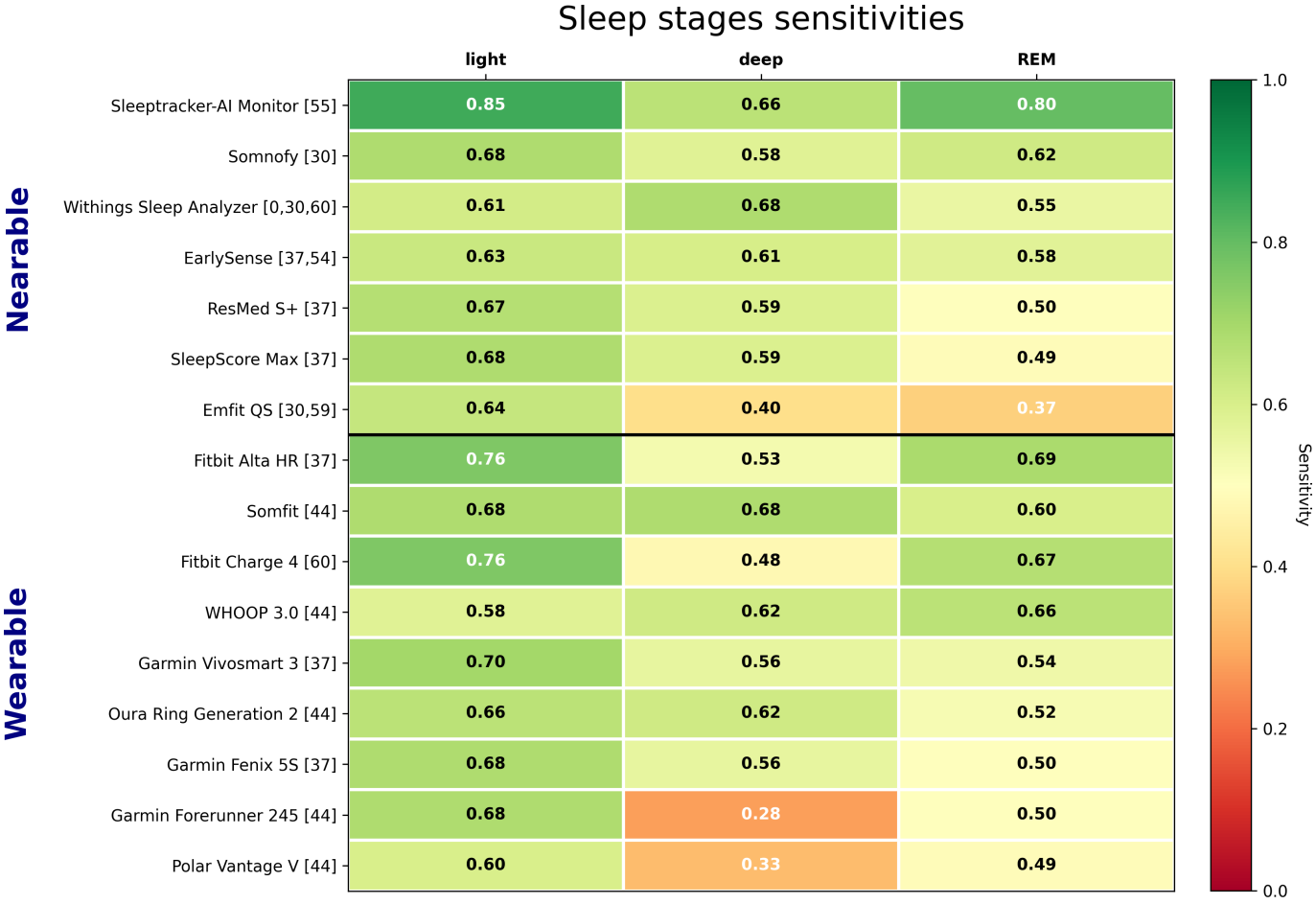
As explained above, the proportion of wake stages in the dataset before and after sleep onset can significantly impact the sleep/wake endpoints. These studies vary in the proportions of wake epochs, which refer to the total time spent awake divided by the total recording duration of the entire dataset. Specifically, these proportions ranged from 8% to 14% in [37], 15% in [54], ranged from 20% to 33% in [44], and from 34% to 42% in [33], compared to 30% in our dataset. This information was unfortunately not available for most of the studies evaluating nearable devices [55-60]. Lower proportions of wake data can inflate overall accuracies and kappas if sleep epoch classification dominates, as noted in [61]. In datasets with a low proportion of wake epochs, overall accuracy and kappa can be misleadingly high

due to the dominance of sleep epoch classification, which is better classified than wake epochs as indicated in the previous paragraph.

Sleep Stages Classification

The trade-off observed in sleep-wake classification is even more pronounced in sleep stage analysis, where no single device consistently outperforms the others across all stages (Figure 14). The range of sensitivities is particularly large for REM, from 0.37 to 0.80, and quite large for deep sleep (0.28 to 0.68), even after excluding wearables (0.40 to 0.68). The sensitivities for light sleep detection are more homogeneous, with 12 sensitivities between 0.6 and 0.7. The WSA is in the top tier of devices for deep sleep classification, but midrange for light sleep and REM.

Figure 14. Overview of sleep stage sensitivities from varying validation studies. Indication [0] denotes this study. Note: Direct comparison is limited by differences in study protocols and populations [33,37,44,54,55,59,60].



The confusion matrices show that, for all devices, the highest proportion of misclassification for each stage is wrong light sleep detection, which can be explained by the inherent difficulty in distinguishing light from the other sleep stages by human experts, as shown in the Results section.

Device Comparison Conclusion

This analysis highlights the necessity of considering a comprehensive set of metrics to accurately evaluate the capabilities of each device. Relying on isolated overall metrics such as accuracy or kappa is insufficient to provide a fair assessment [62,63]. Every device exhibits relative strengths and weaknesses, and no single device clearly outperforms the others across all tasks. The WSA specifically demonstrates robust performance in sleep-wake distinction, while the precise identification of individual sleep stages remains a more challenging task.

Methodological Aspects and Strengths of the Study

The rapid adoption of new sleep technologies in consumer and clinical settings demands rigorous performance evaluation [64]. To improve the reliability and comparability of sleep research, we describe our methodology in detail and with transparency, an approach that is supported by recent guidelines advocating for uniform analysis pipelines [41,42,44]. Our study was designed to resolve methodological issues frequently found in the literature.

The study's strength lies first in its diverse sample of over 100 participants, who are for the most part healthy and with diverse demographic characteristics. Additionally, the inclusion of relevant factors that may negatively impact the device's performance, like the presence of a bed partner and variations in mattress type and thickness, further enriches

the dataset, allowing for a comprehensive evaluation of the device's robustness across different real-life conditions in a healthy population. This allowed us to conduct meaningful subgroup analyses.

A second strength of the study is the method of PSG analysis. Unlike [44,55-58], which opted for a single reviewer, or two reviewers without consensus as in [37], our study is aligned with [33,38] by implementing a double review process with adjudication by a third reviewer in case of disagreement. This consensus methodology is preferable since it enhances the reliability of the sleep stage classification. In particular, we have shown that reviewers often err by misclassifying light sleep stages, a finding consistent with previous research reporting that the interrater reliabilities for light and deep stages were moderate [61]. A reliable assessment of devices' performance in sleep stage classification therefore requires a strong methodology to establish the ground truth.

Third, with a view to limit the participants' discomfort, we chose to perform PSG at home with a limited number of sensors, without the nasal cannula and oximeter. This setup reduces the sleep disturbances typically associated with lab-based PSG, allowing for a more natural assessment of sleep patterns. This limited set of sensors is sufficient for the evaluation of TST and sleep stage classification, with the trade-off that sleep onset, time in bed, and wakefulness after sleep onset cannot be assessed. Yet, significant sleep disturbances were reported by the participant during the PSG test, underscoring the challenges of completely undisturbed monitoring even in home environments. Despite this, the study offers a realistic evaluation of the WSA's capabilities under typical sleeping conditions.

Fourth, we aimed to give a comprehensive report of the device's performance. Collecting epoch-by-epoch data enables a granular examination of sleep patterns, the consistency of measurement errors over time, and thus enables an accurate assessment of the reliability and utility of sleep monitoring devices in real-world settings. We emphasized the need for such comprehensive reporting and the elements of the protocol that may bias the performance estimates. Thus, transparency requires a presentation of the results including all the performance endpoints, each presenting a different facet of the device's performance [41-43].

In conclusion, given these methodological strengths, our study provides valuable insights into the WSA sleep stage algorithm's performance in real-world settings.

Limitations and Future Work

Despite our efforts for a solid methodology, the study faced several challenges.

Challenges in Defining Bedtime and Wake Time in Sleep Studies

A first challenge is created by the lack of a standardized definition of the start time and end time of a sleep study. As per AASM guidelines [50], bedtime and wake time are defined as the "lights-off" and "lights-on," respectively. This is not adapted to studies in real-world settings, where participants often go to bed and engage in activities before attempting to fall asleep. In everyday life, individuals may not turn off the lights immediately upon going to bed, nor do they always intend to sleep right away. Similarly, in the morning, people may not turn on the lights immediately upon awakening, and their actual wake time may not coincide with the "lights-on" reference. This discrepancy is compounded by the lack of a position sensor, which makes it difficult to accurately determine when a person intends to sleep. In other studies [37], start and end times of sleep are determined by the participants themselves, as suggested in standard sleep diaries [65], but the precision of this method is poor.

This lack of standardization is detrimental to the assessment of nearable and wearable devices intended to be used in home settings. Errors in sleep estimation, where wakefulness is mistakenly classified as sleep, are more frequent at the beginning and end of the night. The reason is that many devices, including the WSA, rely heavily on movement detection to distinguish wake from sleep. Thus, wake periods when the participant is immobile and calm in bed, such as when reading or scrolling with a low heart rate, are frequently mistaken as sleep by the device.

Additionally, the practice of activating the PSG immediately before sleep and deactivating it early prevents the gathering of comprehensive data on a participant's sleep habits. This includes behaviors such as spending extended periods in bed without the intention to sleep. To address this limitation, future studies could extend the duration of PSG recordings to capture activities that occur between the time a person gets into bed and falls asleep, as well as between waking and getting out of bed. This approach would provide a more complete picture of sleep behaviors and enable a thorough evaluation of sleep stage classification algorithms in real-world settings.

Challenges in Data Collection From a Nearable

A second challenge in studies with connected devices is data loss. Our study is no exception since nearly 19% of the data were lost, primarily due to the wireless transfer by WSA of a large amount of raw sensor data (250 samples per second). Note that this high rate is not representative of the intended use of the device, which, in its commercial version, only transfers processed data at a low frequency. To mitigate the risk of data loss, Wi-Fi routers paired with the WSA were installed at the participant's home. The routers experienced occasional connectivity problems. Data loss was further exacerbated by issues such as accidental unplugging of the device and improper reconnection without calibration. These technical challenges underscore the difficulties of gathering high-quality, high-frequency data in uncontrolled settings.

Another factor impacted the accuracy of sleep stage classification; despite receiving instructions from a technician, participants with large beds did not always sleep above the sensor but on the other side of the bed. In such cases, the

WSA is not operated optimally, which may have restricted its performance.

Future Work

To validate the utility of longitudinal passive data collection, it is essential to demonstrate that errors in sleep quality measures are consistent over time for a given user.

In addition, our study population only includes healthy participants. There is no data on the performance of the device in special populations that might benefit the most from a longitudinal follow-up.

Looking ahead, the contactless devices hold significant potential for a wide range of research and clinical applications [14,64,66-70]. Their capability to monitor sleep in natural environments [68] underscores their versatility and relevance in advancing our understanding of sleep health and its implications.

Conclusions

In this study, we aimed to (1) assess the performance of a contactless device for sleep stage identification against the gold standard PSG, (2) benchmark it against other comparable devices, and (3) better understand the limitations of such devices. Our paper contributes to those, while also laying out important methodological issues.

This study contributes to the growing body of evidence supporting the utility of nearables for longitudinal sleep monitoring. The WSA's contactless, undermattress design facilitates unobtrusive data collection in free-living conditions and demonstrates sleep-wake detection performance comparable to established consumer devices. While precise sleep stage classification remains a technical challenge, our results highlight the importance of testing on large, diverse, challenging datasets to capture real-world variability. This research underscores the need for standardized performance assessment methodologies to enable reliable comparisons across devices and studies, offering insights for future device design and guidance for utilization, advancing the goal of improved sleep monitoring for the general population [71].

Acknowledgments

We would like to acknowledge Paul Edouard and Sébastien Ducrot, both employees of Withings, for their crucial support and technical expertise in the offline run of the algorithm. We are also grateful to two anonymous reviewers and the journal editor for their suggestions and comments, which contributed to improving this manuscript. We declare the use of generative AI in the writing process; according to the GAIDeT taxonomy (2025), proofreading was delegated to the Gemini Pro 2.5 tool under full human supervision.

Funding

This study received funding from Withings.

Data Availability

The dataset analyzed during this study is not publicly available due to concerns for privacy, ethical, and legal reasons, but queries for data access can be directed to the corresponding author at privacy@withings.com.

Authors' Contributions

Conceptualization: MAS
Methodology: MAS (lead), DC (lead), MDG (equal), GDL (equal)
Software: MAS
Formal analysis: MAS
Investigation: MAS
Data Curation: MAS
Visualization: MAS
Writing – original draft preparation: MAS
Supervision: DC (lead), MDG (lead), GDL (lead), PAG (equal), ASI (equal)
Writing – review & editing: DC (lead), MDG (equal), GDL (equal), PAG (supporting), ASI (supporting)

Conflicts of Interest

This study received funding from Withings, the manufacturer of the device evaluated in this study. MAS is a PhD student affiliated with Université Paris Cité, Reykjavík University, and Withings. MDG, GDL, and DC are employees of Withings. PE is a consultant for Withings. Withings employees contributed to all the stages of this study, from conceptualization to manuscript writing. No other competing interests.

Multimedia Appendix 1

Details on methods and algorithm performance per sleep stage.
[DOCX File (Microsoft Word File), 3319 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Results of Kruskal-Wallis and Brown-Forsythe statistical tests followed by Holm-Bonferroni correction.
[XLSX File (Microsoft Excel File), 39 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Comparison with previous research.
[XLSX File (Microsoft Excel File), 66 KB-Multimedia Appendix 3]

References

1. Steptoe A, O'Donnell K, Marmot M, Wardle J. Positive affect, psychological well-being, and good sleep. J Psychosom Res. Apr 2008;64(4):409-415. [doi: 10.1016/j.jpsychores.2007.11.008]
2. O'Leary K, Bylsma LM, Rottenberg J. Why might poor sleep quality lead to depression? A role for emotion regulation. Cogn Emot. Dec 2017;31(8):1698-1706. [doi: 10.1080/02699931.2016.1247035] [Medline: 27807996]
3. Lallukka T, Sivertsen B, Kronholm E, Bin YS, Øverland S, Glozier N. Association of sleep duration and sleep quality with the physical, social, and emotional functioning among Australian adults. Sleep Health. Apr 2018;4(2):194-200. [doi: 10.1016/j.sleh.2017.11.006]
4. Grandner MA. Sleep, health, and society. Sleep Med Clin. Jun 2022;17(2):117-139. [doi: 10.1016/j.jsmc.2022.03.001] [Medline: 35659068]
5. Geoffroy PA, Hoertel N, Etain B, et al. Insomnia and hypersomnia in major depressive episode: prevalence, sociodemographic characteristics and psychiatric comorbidity in a population-based study. J Affect Disord. Jan 2018;226:132-141. [doi: 10.1016/j.jad.2017.09.032]
6. Morin CM, Jarrin DC. Epidemiology of insomnia: prevalence, course, risk factors, and public health burden. Sleep Med Clin. 2013;8(3). [doi: 10.1016/j.jsmc.2013.05.002] [Medline: 35659072]
7. Young TB. Epidemiology of daytime sleepiness: definitions, symptomatology, and prevalence. J Clin Psychiatry. 2004;65 Suppl 16(Suppl 16):12-16. [Medline: 15575799]
8. Senaratna CV, Perret JL, Lodge CJ, et al. Prevalence of obstructive sleep apnea in the general population: a systematic review. Sleep Med Rev. Aug 2017;34:70-81. [doi: 10.1016/j.smrv.2016.07.002]
9. Rundo JV, Downey RI. Polysomnography. Handb Clin Neurol. 2019;160:381-392. [doi: 10.1016/B978-0-444-64032-1.00025-4] [Medline: 31277862]
10. Rajaratnam SMW, Barger LK, Lockley SW, et al. Sleep disorders, health, and safety in police officers. JAMA. Dec 21, 2011;306(23):2567-2578. [doi: 10.1001/jama.2011.1851]
11. Bloch KE. Polysomnography: a systematic review. Technol Health Care. Oct 1997;5(4):285-305. [doi: 10.3233/THC-1997-5403] [Medline: 9429270]
12. Keenan SA. Polysomnographic technique: an overview. In: Chokroverty S, editor. Sleep Disorders Medicine: Basic Science, Technical Considerations, and Clinical Aspects. Butterworth-Heinemann; 1994:79-94. [doi: 10.1016/B978-0-7506-9002-7.50011-2]
13. Kang SG, Kang JM, Ko KP, Park SC, Mariani S, Weng J. Validity of a commercial wearable sleep tracker in adult insomnia disorder patients and good sleepers. J Psychosom Res. Jun 2017;97:38-44. [doi: 10.1016/j.jpsychores.2017.03.009]
14. Lechat B, Scott H, Manners J, et al. Multi-night measurement for diagnosis and simplified monitoring of obstructive sleep apnoea. Sleep Med Rev. Dec 2023;72:101843. [doi: 10.1016/j.smrv.2023.101843]
15. Toussaint M, Luthringer R, Schaltenbrand N, et al. First-night effect in normal subjects and psychiatric inpatients. Sleep. Jul 1995;18(6):463-469. [doi: 10.1093/sleep/18.6.463] [Medline: 7481418]
16. Smith MT, McCrae CS, Cheung J, et al. Use of actigraphy for the evaluation of sleep disorders and circadian rhythm sleep-wake disorders: an American Academy of Sleep Medicine systematic review, meta-analysis, and GRADE assessment. J Clin Sleep Med. Jul 15, 2018;14(07):1209-1230. [doi: 10.5664/jcsm.7228]
17. Kapur VK, Auckley DH, Chowdhuri S, et al. Clinical practice guideline for diagnostic testing for adult obstructive sleep apnea: an American Academy of Sleep Medicine clinical practice guideline. J Clin Sleep Med. Mar 15, 2017;13(03):479-504. [doi: 10.5664/jcsm.6506]
18. Braun M, Stockhoff M, Tijssen M, Dietz-Terjung S, Coughlin S, Schöbel C. A systematic review on the technical feasibility of home-polysomnography for diagnosis of sleep disorders in adults. Curr Sleep Medicine Rep. Jun 1, 2024;10(2):276-288. [doi: 10.1007/s40675-024-00301-z]
19. Mykytyn IJ, Sajkov D, Neill AM, Mc Evoy RD. Portable computerized polysomnography in attended and unattended settings. Chest. Jan 1999;115(1):114-122. [doi: 10.1378/chest.115.1.114]

20. Andrade L, Paiva T. Ambulatory versus laboratory polysomnography in obstructive sleep apnea: comparative assessment of quality, clinical efficacy, treatment compliance, and quality of life. J Clin Sleep Med. Aug 15, 2018;14(8):1323-1331. [doi: 10.5664/jcsm.7264] [Medline: 30092885]
21. Van de Water ATM, Holmes A, Hurley DA. Objective measurements of sleep for non-laboratory settings as alternatives to polysomnography – a systematic review. J Sleep Res. Mar 2011;20(1 Pt 2):183-200. [doi: 10.1111/j.1365-2869.2009.00814.x] [Medline: 20374444]
22. Pan Q, Brulin D, Campo E. Current status and future challenges of sleep monitoring systems: systematic review. JMIR Biomed Eng. Aug 26, 2020;5(1):e20921. [doi: 10.2196/20921]
23. Vitazkova D, Kosnacova H, Turonova D, et al. Transforming sleep monitoring: review of wearable and remote devices advancing home polysomnography and their role in predicting neurological disorders. Biosensors (Basel). Feb 17, 2025;15(2):117. [doi: 10.3390/bios15020117] [Medline: 39997019]
24. Sivertsen B, Omvik S, Havik OE, et al. A comparison of actigraphy and polysomnography in older adults treated for chronic primary insomnia. Sleep. Oct 2006;29(10):1353-1358. [doi: 10.1093/sleep/29.10.1353]
25. Paquet J, Kawinska A, Carrier J. Wake detection capacity of actigraphy during sleep. Sleep. Oct 2007;30(10):1362-1369. [doi: 10.1093/sleep/30.10.1362] [Medline: 17969470]
26. de Zambotti M, Goldstein C, Cook J, et al. State of the science and recommendations for using wearable technology in sleep and circadian research. Sleep. Apr 12, 2024;47(4):38149978. [doi: 10.1093/sleep/zsad325]
27. Kolla BP, Mansukhani S, Mansukhani MP. Consumer sleep tracking devices: a review of mechanisms, validity and utility. Expert Rev Med Devices. May 2016;13(5):497-506. [doi: 10.1586/17434440.2016.1171708] [Medline: 27043070]
28. Edouard P, Campo D, Bartet P, et al. Validation of the Withings Sleep Analyzer, an under-the-mattress device for the detection of moderate-severe sleep apnea syndrome. J Clin Sleep Med. Jun 2021;17(6):1217-1227. [doi: 10.5664/jcsm.9168]
29. Lechat B, Naik G, Reynolds A, et al. Multinight prevalence, variability, and diagnostic misclassification of obstructive sleep apnea. Am J Respir Crit Care Med. Mar 1, 2022;205(5):563-569. [doi: 10.1164/rccm.202107-1761OC]
30. G Ravindran KK, della Monica C, Atzori G, et al. Reliable contactless monitoring of heart rate, breathing rate, and breathing disturbance during sleep in aging: digital health technology evaluation study.JMIR Mhealth Uhealth. 2024;12:e53643. [doi: 10.2196/53643]
31. Jagielski JT, Bibi N, Gay PC, et al. Evaluating an under-mattress sleep monitor compared to a peripheral arterial tonometry home sleep apnea test device in the diagnosis of obstructive sleep apnea. Sleep Breath. Aug 2023;27(4):1433-1441. [doi: 10.1007/s11325-022-02751-7]
32. Huyett P, Yu PK, Kent D, et al. Comparison of home sleep devices and sleep study testing in hypoglossal nerve stimulation patients. Laryngoscope. Jan 2025;135(1):469-477. [doi: 10.1002/lary.31710] [Medline: 39162306]
33. Ravindran KKG, della Monica C, Atzori G, et al. Contactless and longitudinal monitoring of nocturnal sleep and daytime naps in older men and women: a digital health technology evaluation study. Sleep. Oct 11, 2023;46(10):zsad194. [doi: 10.1093/sleep/zsad194]
34. Yang RY, Bendjoudi A, Buard N, Boutouyrie P. Pneumatic sensor for cardiorespiratory monitoring during sleep. Biomed Phys Eng Express. 2019;5(5):055014. [doi: 10.1088/2057-1976/ab3ac9]
35. Hedner J, White DP, Malhotra A, et al. Sleep staging based on autonomic signals: a multi-center validation study. J Clin Sleep Med. Jun 15, 2011;7(3):301-306. [doi: 10.5664/JCSM.1078] [Medline: 21677901]
36. Toften S, Pallesen S, Hrozanova M, Moen F, Grønli J. Validation of sleep stage classification using non-contact radar technology and machine learning (Somnofy®). Sleep Med. Nov 2020;75:54-61. [doi: 10.1016/j.sleep.2020.02.022]
37. Chinoy ED, Cuellar JA, Huwa KE, et al. Performance of seven consumer sleep-tracking devices compared with polysomnography. Sleep. May 14, 2021;44(5):zsaa291. [doi: 10.1093/sleep/zsaa291] [Medline: 33378539]
38. Kainec KA, Caccavaro J, Barnes M, Hoff C, Berlin A, Spencer RMC. Evaluating accuracy in five commercial sleep-tracking devices compared to research-grade actigraphy and polysomnography. Sensors (Basel). 2024;24(2):635. [doi: 10.3390/s24020635]
39. Nelson BW, Saeb S, Barman P, et al. Performance evaluation of the verily numetric watch sleep suite for digital sleep assessment against in-lab polysomnography. J Sleep Res. Oct 2025;34(5):e70036. [doi: 10.1111/jsr.70036] [Medline: 40170300]
40. Willoughby AR, Golkashani HA, Ghorbani S, et al. Performance of wearable sleep trackers during nocturnal sleep and periods of simulated real-world smartphone use. Sleep Health. Jun 2024;10(3):356-368. [doi: 10.1016/j.sleh.2024.02.007] [Medline: 38570223]
41. Benedetti D, Menghini L, Vallat R, et al. Call to action: an open-source pipeline for standardized performance evaluation of sleep-tracking technology. Sleep. Feb 8, 2023;46(2):zsac304. [doi: 10.1093/sleep/zsac304] [Medline: 36611112]

42. Menghini L, Cellini N, Goldstone A, Baker FC, de Zambotti M. A standardized framework for testing the performance of sleep-tracking technology: step-by-step guidelines and open-source code. Sleep. Feb 12, 2021;44(2):zsaa170. [doi: 10.1093/sleep/zsaa170] [Medline: 32882005]
43. Chee MWL, Baumert M, Scott H, et al. World Sleep Society recommendations for the use of wearable consumer health trackers that monitor sleep. Sleep Med. Jul 2025;131:106506. [doi: 10.1016/j.sleep.2025.106506]
44. Miller DJ, Sargent C, Roach GD. A validation of six wearable devices for estimating sleep, heart rate and heart rate variability in healthy adults. Sensors (Basel). Aug 22, 2022;22(16):6317. [doi: 10.3390/s22166317] [Medline: 36016077]
45. Birrer V, Elgendi M, Lambercy O, Menon C. Evaluating reliability in wearable devices for sleep staging. NPJ Digit Med. Mar 18, 2024;7(1):74. [doi: 10.1038/s41746-024-01016-9] [Medline: 38499793]
46. Yuan H, Hill EA, Kyle SD, Doherty A. A systematic review of the performance of actigraphy in measuring sleep stages. J Sleep Res. Oct 2024;33(5):e14143. [doi: 10.1111/jsr.14143] [Medline: 38384163]
47. Imtiaz SA. A systematic review of sensing technologies for wearable sleep staging. Sensors (Basel). Jan 2021;21(5):1562. [doi: 10.3390/s21051562]
48. Berry RB. Fundamentals of Sleep Medicine. Elsevier; 2012. ISBN: 9781437703269
49. Buysse DJ, Reynolds CF III, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. Psychiatry Res. May 1989;28(2):193-213. [doi: 10.1016/0165-1781(89)90047-4] [Medline: 2748771]
50. Berry RB, Brooks R, Gamaldo CE, Harding SM, Marcus CL, Vaughn BV. The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications. American Academy of Sleep Medicine; 2015. URL: https://aasm.org/resources/pdf/scoring-manual-preface.pdf [Accessed 2026-02-14]
51. Schmidt M, Schumann A, Müller J, Bär KJ, Rose G. ECG derived respiration: comparison of time-domain approaches and application to altered breathing patterns of patients with schizophrenia. Physiol Meas. Apr 1, 2017;38(4):601-615. [doi: 10.1088/1361-6579/aa5feb]
52. de Zambotti M, Goldstone A, Claudatos S, Colrain IM, Baker FC. A validation study of Fitbit Charge 2™ compared with polysomnography in adults. Chronobiol Int. Apr 2018;35(4):465-476. [doi: 10.1080/07420528.2017.1413578] [Medline: 29235907]
53. Lee YJ, Lee JY, Cho JH, Choi JH. Interrater reliability of sleep stage scoring: a meta-analysis. J Clin Sleep Med. Jan 2022;18(1):193-202. [doi: 10.5664/jcsm.9538]
54. Tal A, Shinar Z, Shaki D, Codish S, Goldbart A. Validation of contact-free sleep monitoring device with comparison to polysomnography.J Clin Sleep Med. Mar 15, 2017;13(3):517-522. [doi: 10.5664/jcsm.6514] [Medline: 27998378]
55. Ding F, Cotton-Clay A, Fava L, et al. Polysomnographic validation of an under-mattress monitoring device in estimating sleep architecture and obstructive sleep apnea in adults. Sleep Med. Aug 2022;96:20-27. [doi: 10.1016/j.sleep.2022.04.010]
56. Ellender CM, Zahir SF, Meaklim H, Joyce R, Cunnington D, Swieca J. Prospective cohort study to evaluate the accuracy of sleep measurement by consumer-grade smart devices compared with polysomnography in a sleep disorders population. BMJ Open. Nov 2021;11:e044015. [doi: 10.1136/bmjopen-2020-044015]
57. Hsiou DA, Gao C, Matlock RC, Scullin MK. Validation of a nonwearable device in healthy adults with normal and short sleep durations. J Clin Sleep Med. Mar 1, 2022;18(3):751-757. [doi: 10.5664/jcsm.9700] [Medline: 34608858]
58. Tuominen J, Peltola K, Saaresranta T, Valli K. Sleep parameter assessment accuracy of a consumer home sleep monitoring ballistocardiograph beddit sleep tracker: a validation study.J Clin Sleep Med. Mar 15, 2019;15(3):483-487. [doi: 10.5664/jcsm.7682] [Medline: 30853052]
59. Kholghi M, Szollosi I, Hollamby M, Bradford D, Zhang Q. A validation study of a ballistocardiograph sleep tracker against polysomnography. J Clin Sleep Med. Apr 1, 2022;18(4):1203-1210. [doi: 10.5664/jcsm.9754] [Medline: 34705630]
60. Manners J, Kemps E, Lechat B, Catcheside P, Eckert DJ, Scott H. Performance evaluation of an under-mattress sleep sensor versus polysomnography in > 400 nights with healthy and unhealthy sleep. J Sleep Res. Dec 2025;34(6):e14480. [doi: 10.1111/jsr.14480] [Medline: 40017337]
61. Kelly JM, Strecker RE, Bianchi MT. Recent developments in home sleep-monitoring devices. ISRN Neurol. Oct 14, 2012;2012:1-10. [doi: 10.5402/2012/768794]
62. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol. May 1993;46(5):423-429. [doi: 10.1016/0895-4356(93)90018-v] [Medline: 8501467]
63. Gjørup T. The kappa coefficient and the prevalence of a diagnosis. Methods Inf Med. Oct 1988;27(4):184-186. [doi: 10.1055/s-0038-1635539] [Medline: 3226333]

64. Pellegrini M, Lannin NA, Mychasiuk R, Graco M, Kramer SF, Giummarra MJ. Measuring sleep quality in the hospital environment with wearable and non-wearable devices in adults with stroke undergoing inpatient rehabilitation.Int J Environ Res Public Health. Feb 23, 2023;20(5):3984. [doi: 10.3390/ijerph20053984] [Medline: 36900995]
65. Carney CE, Buysse DJ, Ancoli-Israel S, et al. The consensus sleep diary: standardizing prospective sleep self-monitoring. Sleep. Feb 2012;35(2):287-302. [doi: 10.5665/sleep.1642]
66. Lechat B, Naik G, Appleton S, et al. Multi-night assessment of snoring is associated with hypertension independent of OSA severity [Abstract]. Am J Respir Crit Care Med. 2023;207:A6200. [doi: 10.1164/ajrccm-conference.2023.207.1_MeetingAbstracts.A6200]
67. de Havenon A, Falcone G, Rivier C, et al. Impact of sleep quality and physical activity on blood pressure variability. PLoS ONE. 2024;19(4):e0301631. [doi: 10.1371/journal.pone.0301631]
68. Scott H, Lechat B, Guyett A, et al. Sleep irregularity is associated with hypertension: findings from over 2 million nights with a large global population sample. Hypertension. May 2023;80(5):1117-1126. [doi: 10.1161/HYPERTENSIONAHA.122.20513]
69. Scott H, Naik G, Lechat B, et al. Are we getting enough sleep? Frequent irregular sleep found in an analysis of over 11 million nights of objective in-home sleep data. Sleep Health. Feb 2024;10(1):91-97. [doi: 10.1016/j.sleh.2023.10.016]
70. Chapple C, Bliwise D, Maislisch L, Roitmann E, Burtea T. Night-time voids, level of bother and sleep characteristics in a non-patient population of wearable devices users. Int J Clin Pract. Jul 2020;74(7):e13495. [doi: 10.1111/ijcp.13495] [Medline: 32100396]
71. Chaudhry FF, Danieletto M, Golden E, et al. Sleep in the natural environment: a pilot study. Sensors (Basel). 2020;20(5):1378. [doi: 10.3390/s20051378]

Abbreviations

AASM: American Academy of Sleep Medicine
ECG: electrocardiogram
EEG: electroencephalogram
EMG: electromyogram
EOG: electrooculogram
MAE: mean absolute error
MCAR: missing completely at random
NEDeep: number of episodes in deep sleep stage
NELight: number of episodes in light sleep stage
NEREM: number of episodes in REM sleep stage
NEWake: number of episodes in wake stage
PIDeep: proportion of the night in deep sleep stage
PILight: proportion of the night in light sleep stage
PIREM: proportion of the night in REM sleep stage
PIWake: proportion of the night in wake stage
PSG: polysomnography
PSQI: Pittsburgh Sleep Quality Index
REM: rapid eye movement (sleep stage)
TIDeep: time spent in deep sleep stage
TILight: time spent in light sleep stage
TIREM: time spent in REM sleep stage
TIWake: time spent in wake stage
TST: total sleep time
WSA: Withings Sleep Analyzer

Edited by James Waterson; peer-reviewed by Benjamin W Nelson, F A Etindele Sosso, Lucia Pinilla; submitted 06.May.2025; final revised version received 10.Jan.2026; accepted 26.Jan.2026; published 02.Apr.2026

Please cite as:
Stefanos MA, de Laboulaye G, Campo D, de Gourcuff M, Escourrou P, Matrot B, Islind AS, Geoffroy PA
Evaluation of a Contactless Sleep Monitoring Device for Sleep Stage Detection at Home in a Healthy Population: Prospective Study in Free-Living Conditions
JMIR Hum Factors 2026;13:e77033
URL: https://humanfactors.jmir.org/2026/1/e77033
doi: 10.2196/77033

© Marie-Ange Stefanos, Guillaume de Laboulaye, David Campo, Martin de Gourcuff, Pierre Escourrou, Boris Matrot, Anna Sigridur Islind, Pierre Alexis Geoffroy. Originally published in JMIR Human Factors (https://humanfactors.jmir.org), 02.Apr.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Human Factors, is properly cited. The complete bibliographic information, a link to the original publication on https://humanfactors.jmir.org, as well as this copyright and license information must be included.

3.3 Chapter Summary

- Prospective at-home evaluation of the Withings Sleep Analyzer sleep-staging algorithm against simultaneous home PSG in 117 healthy adults (69 women; mean age 39.9 years, SD 11.4), free-living, assessing agreement with the reference for both sleep-wake distinction and sleep-stage identification.
- Sleep-wake distinction: 87% overall accuracy (95% CI 86–87%), high sleep sensitivity (93%, 95% CI 92–94%), moderate wakefulness sensitivity (73%, 95% CI 69–77%).
- Sleep-stage identification: mean accuracy 63% (95% CI 62–65%), Cohen κ 0.49 (95% CI 0.47–0.51), most errors falling into light sleep (per-stage accuracy REM 53%, light 68%, deep 56%).
- Systematic biases: overestimated total sleep time (+20 min) and light sleep (+1 h 21 min); underestimated REM (−15 min) and deep sleep (−46 min).
- Robustness: consistent across age, sex, BMI, self-reported sleep quality and PSQI, mattress thickness, and bed-partner presence, the only exception being wake proportion across subjective sleep-duration groups.
- Performance was measured over the full in-bed period rather than between AASM lights-off and lights-on, because the proportion of pre- and post-sleep wakefulness inflates wake-detection metrics; this gives a more conservative, unbiased estimate.
- The device's stage errors fell mainly into light sleep and mirrored disagreements between expert reviewers, who disagree most on light versus other stages, so accuracy is bounded by the inter-rater variability of the training labels, an imperfect gold standard.
- The free-living design captures real-world diversity and improves generalizability but is a harder scenario, since reading or screen time in bed can be confounded with sleep, giving a more realistic estimate than controlled-laboratory studies.
- Against published consumer-device validations, the device sits in the higher-specificity tier, with slightly lower sensitivity comparable to EarlySense in contrast to high-sensitivity low-specificity wearables, ranks top-tier for deep sleep and midrange for light and REM, light sleep being the hardest stage for all devices.
- The device did not degrade perceived sleep, whereas the PSG night itself altered participants' perceived sleep quality.
- Take-home: a promising contactless option for long-term home monitoring with competitive performance, but accuracy gains and a standardized evaluation methodology are needed, particularly for populations with impaired sleep.

Chapter 4

Data Collection: SLEEPI Protocol

4.1 Introduction

Developing machine learning models that detect insomnia from physiological data requires large, longitudinal datasets recorded under real-world conditions, which traditional site-based clinical studies struggle to produce given their logistical burden and limited scale. Decentralized clinical trials conducted through consumer devices offer a way to reach this scale, but they introduce new difficulties in recruiting and retaining participants and in managing the data they generate. This chapter presents the Smart devices-based Longitudinal Assessment of Patterns of Insomnia (SLEEPI) study, a large-scale decentralized protocol that collects two paired data streams from participants over a 26-week period in free-living conditions: subjective reports anchored on clinically validated insomnia criteria, and objective physiological measurements captured by unobtrusive Withings wearable and nearable sensors. The protocol automates the entire participant journey from recruitment to withdrawal with minimal manual intervention, and sizes recruitment by simulating the minimal sample required for the machine learning objective and then applying a prospective attrition model. The resulting dataset is intended to train the insomnia-detection models and provides the data foundation for the analyses presented in the next chapter. The discussion then examines the trade-offs and selection biases of consumer-device data and the infrastructure demands of decentralized trials, and positions the protocol as a template transferable to other health domains.

4.2 Paper

Original Paper:

End-to-end Protocol for large-scale, digital, longitudinal and decentralized clinical trials: the SLEEPI study on insomnia patterns

Abstract [389 words]

Background: Personalized medicine requires understanding health conditions in real-world settings, but traditional clinical tools are often poorly suited for longitudinal tracking due to logistical hurdles, high participant burden, and data quality issues. While consumer wearables offer a scalable alternative, decentralized clinical trials (DCTs) introduce new challenges in patient recruitment, sustained engagement, and data management. Insomnia serves as a suitable use case for this protocol due to its global prevalence, frequent underdiagnosis, and the well-documented discrepancy between subjective patient reports and objective physiological metrics. By providing a large-scale dataset, this protocol addresses these limitations to better reflect the real-world conditions of patient monitoring than using traditional assessment methods.

Objectives: We present and validate a novel research protocol for large-scale DCTs, applied to a longitudinal study on insomnia called the SLEEPI study, by leveraging an established commercial ecosystem with limited additional manual intervention.

Methods: Two data streams are collected: subjective, clinically-anchored patient reports, and objective physiological metrics (e.g., heart and respiratory rates) captured via unobtrusive wearable and nearable sensors. This study is executed through a highly automated, end-to-end digital protocol designed to conduct a large-scale, longitudinal clinical trial. This protocol deliberately reduces the resource-intensive, manual interventions of traditional studies with a suite of digital tools and strategies that manage the entire participant journey. We detail this highly automated framework, from initial recruitment to engagement, adherence, study monitoring, data management, and withdrawal. The data collected are intended to train a machine learning model. The recruitment objectives were determined by using simulations to determine the minimal sample size for that purpose, and by applying a prospective attrition model to calculate the initial contact volume needed to reach that targeted sample.

Results: This protocol enabled the continuous monitoring of over 10,000 individuals for 26 weeks, creating the largest and most diverse longitudinal insomnia dataset to date. Enrolment metrics significantly exceeded projections: initial survey response rates reached 16% (vs. 10% projected), the consent rate was 96% (vs. 77% projected), and the final inclusion rate was 90% (vs. 81% projected), therefore validating the recruitment efficiency and scalability of this protocol.

Conclusions: This framework reduces operational demands by leveraging pre-existing user bases of connected devices. Beyond insomnia, the protocol provides a blueprint for generating large-scale, real-world clinical datasets, presenting new opportunities for research in other complex areas such as immunology, mental health, and cardiovascular diseases.

Trial Registration: ClinicalTrials.gov NCT07430878;
https://clinicaltrials.gov/study/NCT07430878

INTRODUCTION

Conducting large-scale, decentralized, and longitudinal clinical studies in real-world settings is essential for better understanding health conditions[1], yet such research faces significant logistical and methodological challenges, including recruitment bottlenecks, participant burden, and data quality assurance[2]. To address these issues, and building on previous decentralized trials models[3–5], we propose a new, scalable protocol that harnesses an established commercial smart device ecosystem. The protocol includes several innovations designed to dramatically reduce operational demands: leveraging a pre-existing user base, passive data collection, automated data management, structured withdrawal process, and a model to forecast participant dropout at each stage of the study, which enables accurate resource planning.

We developed this new methodology for a decentralized clinical trial on insomnia, a field long in need of innovative methods. Insomnia is a prevalent chronic sleep disorder characterized by difficulties in initiating or maintaining sleep, leading to significant daytime impairments such as fatigue and concentration difficulties[6,7]. The disorder affects approximately 30% of the world population[8–10], yet it remains frequently underdiagnosed, with fewer than 15% of affected individuals seeking medical intervention[11]. A challenge in its characterization is the notable discrepancy between patients' subjective reports and objective metrics from standard assessment tools[12,13]. Although diagnosis currently relies on clinical evaluation of subjective patient reports, it could be augmented and refined by objective, longitudinal data analysis[11,14]. Our central hypothesis is that machine learning models, when trained on high-dimensional and longitudinal sensor data, can identify objective digital biomarkers predictive of a patient's subjective experience of insomnia, thereby elucidating complex, non-linear relationships that traditional methods like actigraphy fail to capture. In other words, we aim to investigate the possibility of creating insomnia phenotypes with machine learning models of longitudinal sensor data.

The traditional data collection devices have well-documented limitations that impede the application of data science techniques to the clinical research on insomnia : actigraphy can erroneously interpret prolonged wakeful immobility as sleep[15,16], while polysomnography (PSG) is resource-intensive, obtrusive, and generally limited to one or two nights, an inadequate timeframe for capturing the natural variability of sleep patterns over time[17,18]. As a result of these limitations, the dataset collected with these instruments are not only inadequate to train machine learning models, but also present methodological issues. Despite European guidelines recommending large sample sizes[6], studies frequently use up to 100 participants[19,20], which is insufficient to train machine learning (ML) models that can generalize across diverse demographic and clinical subgroups within the study sample, such as sex, age, body mass index (BMI), race, and socioeconomic status[21]. Besides, the data collection periods are also often too brief, typically lasting 1–8 nights[22–27], which falls short of the 14 nights to 2 months recommended to capture the disorder's variability[28]. Furthermore, these datasets usually lack demographic diversity (e.g., male-only cohorts[29]) and rely on data from artificial laboratory settings[30,31] that fail to reflect real-world patient conditions.

To address these clinical and research gaps, we designed the Smart devices-based Longitudinal Assessment of Patterns of Insomnia (SLEEPI) study. This large-scale, decentralized clinical trial

captures real-world data by monitoring each participant in their home with connected wearable and nearable sensors for a 26-week period in order to overcome the limitations of prior research by capturing the episodic nature and variability of insomnia characteristics, such as initial versus terminal insomnia. While this type of dataset has inherent biases requiring careful mitigation, it has the benefit of capturing the demographic and clinical heterogeneity characteristic of real-world populations, which is often absent in the highly-selected cohorts of traditional laboratory studies.

To test our hypothesis that machine learning models can identify new digital biomarkers of insomnia, our methodology captures two synchronized data streams: first, subjective data in the form of clinically-anchored patient-reported outcomes (PROs) from a mobile app, which serve as the ground truth; and second, a rich collection of objective data from Withings' wearable and nearable devices. Previous decentralized clinical trials have demonstrated the feasibility of using consumer-grade devices to create extensive datasets[3,4,32] from which phenotyping of complex disorders into homogeneous subgroups was possible[18,33–36]. In our study, the devices' embedded sensors — photoplethysmography, accelerometers, microphones, and pressure and temperature sensors — enable the passive, unobtrusive, and continuous collection of vital parameters such as heart rate, respiratory rate, and variables related to activity and sleep[28,37]. By applying supervised machine learning analyses to this dataset, we will be able to test the central hypothesis that these passive sensor data contain objective patterns that are strongly correlated with a patient's subjective experience of insomnia.

The verification of our central hypothesis through the SLEEPI study provides a foundational step toward integrating these digital health technologies into clinical practice. Such an integration could enhance multiple facets of patient care, from enabling large-scale screening and improving diagnostic accuracy to facilitating longitudinal monitoring and refining patient phenotyping for more personalized treatment. Ultimately, the protocol's applicability extends far beyond insomnia, as its core strength lies in its ability to link objective and subjective data over time. The continuous collection of objective data allows for the assessment of intra-individual variability in physiological and behavioral metrics, while the concomitant collection of subjective data offers insights into symptom fluctuations[18]. This combined approach is essential for numerous other domains, including cardiovascular monitoring, neurodegenerative disease progression, chronic pain management, and mental health, providing a robust framework for the future of real-world digital health research.

Methods

The French Ethical Committee accepted the study on 25-07-2024, with ID number 2024-A01028-39, by CPP Ouest III.

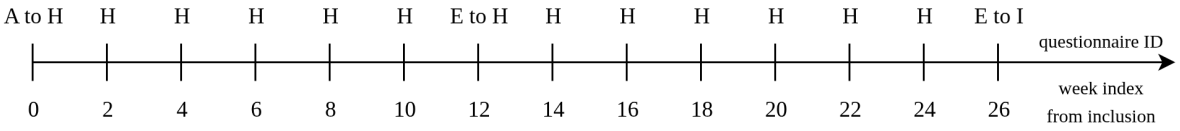
Study Design

The primary objective of the SLEEPI study is to evaluate the ability of an ML model to perform time series binary classification, distinguishing between insomnia and healthy sleepers using

longitudinal, sparse, multivariate data collected in the real environment of the participant and generated by non-obtrusive sensors. The performance of this insomnia detection model will be evaluated using standard classification metrics, such as accuracy, sensitivity, specificity, and confusion matrix. The secondary objective is to assess participant adherence to this longitudinal, decentralized study.

To achieve these goals, the SLEEPI study involves three Withings sleep-related devices: the nearable Withings Sleep Analyzer (WSA) and the wearable smartwatches ScanWatch (SW) and ScanWatch 2 (SW2). Together, these devices continuously record objective data to capture key physiological and behavioral parameters: detailed sleep architecture, activity intensity, temperature variations, and a range of cardiorespiratory variables such as heart rate, heart rate variability (HRV), oxygen saturation, and snoring duration[37–44]. The dataset used to train the ML models exclusively contains physiological biomarkers computed from the raw sensor data. The unprocessed source signals, such as acceleration and photoplethysmography (PPG), are not included. Complementing these objective measurements, subjective data are gathered through digital questionnaires administered via the Withings Mobile Application (WMA). These surveys include questions for insomnia diagnosis based on DSM-5-TR criteria and a suite of clinically validated scales to quantify symptom severity: the Insomnia Severity Index (ISI)[45], the Patient Health Questionnaire-9 (PHQ-9)[46], and the General Anxiety Disorder-7 (GAD-7)[47]. The questionnaires also collect demographic, clinical and environmental data to 1) implement screening procedures to limit low-quality data 2) allow for subgroup analyses 3) identify at-risk populations and 4) adjust for confounding variables.

Following the collection of informed consent and verification that all inclusion criteria are met, each participant begins a 26-week monitoring period under free-living conditions, during which participants are required to use at least one of the study devices (WSA, SW, or SW2) and complete several questionnaires sent via the WMA. The questionnaire timeline is available in Figure 1. Once the monitoring period ends, compensation is provided at 8 weeks thereafter to participants who will have met predefined criteria for data contribution, while individual results are shared with all participants at 16 weeks.



Legend: Questionnaire IDs

A: Eligibility Criteria	D: Demographics, PHQ-9, GAD-7	G: Hypnotics
B: Informed Consent Form	E: Labeling	H: ISI
C: Device-based Screening	F: Comorbidities	I: End-of-study questionnaire

Figure 1: Timing and Sequence of study questionnaire administrations during the monitoring phase
PHQ-9[46]: Patient Health Questionnaire-9
GAD-7[47]: General Anxiety Disorder-7
ISI[45]: Insomnia Severity Index

PRO-Derived Ground Truth for Biomarker Identification

The dataset created has several challenging characteristics: it is large-scale, multidimensional, longitudinal, and inherently sparse, due to highly irregular sampling frequencies that vary at multiple levels: between individuals, across different data dimensions, and even over time for a single variable. This overall complexity renders simple signal processing and rule-based approaches ineffective for reliably identifying insomnia. Therefore, we will adopt a supervised ML approach to automatically discover the predictive patterns within the data. This approach has two objectives: first, a binary classification task to differentiate between individuals with and without insomnia; and second, a regression task to assess the severity of the condition. The inputs for these models will be the passively collected sensor data from the inclusion phase.

For each of these tasks, a supervised ML model requires ground-truth labels for each individual, which serve as the correct answers the model learns to predict from the input data. These labels will be derived from Patient-Reported Outcomes (PRO), obtained through the questionnaires, which are themselves based on standard clinically validated criteria. For distinguishing between individuals with and without insomnia, the label is built on the subjective criteria for insomnia diagnosis from the DSM-5-TR[7]. For evaluating insomnia severity, the severity levels are determined using the Insomnia Severity Index (ISI) questionnaire[45].

Digital Study

This study is executed through a highly automated, end-to-end digital protocol designed to conduct a large-scale, longitudinal clinical trial with minimal staff resources. This protocol deliberately reduces the resource-intensive, manual interventions of traditional studies with a sophisticated suite of digital tools that manage the entire participant journey. The following sections detail each component of this automated framework, from initial recruitment to ongoing engagement and data management.

Recruitment

Eligibility Criteria

Participation in the study is contingent upon the self-verification of several inclusion criteria. Participants must be individuals aged 18 years or older, be eligible for the French social security system, and must not be classified as vulnerable according to the French regulation in force[48]. The collection of informed consent is mandatory. Ownership of at least one Withings sleep monitoring device compatible with the study — SW, SW2, or WSA — is also required, as is the installation of the Withings mobile application, version 5.12 or a more recent version, for the delivery of study questionnaires.

Individuals are excluded if they fail to meet any of these inclusion requirements. Further exclusion criteria, assessed through self-reported data, address potential threats to data integrity. These include sharing the device with another person, device usage below a specified minimum frequency, or the presence of any known device malfunctions, as detailed in Supplementary Materials A.

Communication Channels

Participant recruitment is planned for a 26-week period and is conducted through the Withings Customer Relationship Management (CRM) database and various social media platforms.

Targeting through the CRM database identifies registered Withings users who own and have installed at least one of the sleep monitoring devices required for study inclusion. Eligibility for CRM-based contact requires that these users have demonstrated activity, defined as interaction with either one of the specified sleep devices or the WMA, within the 30-day period prior to outreach. Eligible individuals receive a recruitment email that contains an invitation to participate and includes a button designed to redirect the user into the WMA to directly access the digital forms for eligibility verification and informed consent.

We recruit participants by posting content on social media that was designed to spread naturally through user shares and follows, without using any paid advertising. This content is disseminated through posts and temporarily visible content on selected platforms: LinkedIn, Twitter, Instagram, and Facebook. Instagram account owners who are under marketing partnership contracts with Withings may also contribute to recruitment outreach by sharing this prepared content on a voluntary basis, which is consistent with the study's non-commercial research objectives.

A contingency plan involves targeted, paid advertising via Facebook and Instagram utilizing criteria derived from the internal CRM database. This measure is planned to be implemented only if the number of enrolled participants is below 3,000 individuals after the first month of the recruitment period.

Screening and Consent Process

Informed consent was obtained from all participants involved in the study. The enrollment and consent process is conducted digitally through the WMA. Potential participants first review an electronic information notice, after which they use an integrated digital form to provide informed consent and complete their eligibility verification. Following successful screening and receipt of consent, an automated email confirms the participant's official inclusion in the study and provides a downloadable copy of the information notice for their records.

Timeline

The study was launched in February 2025, with an initial recruitment phase planned for 26 weeks. As each participant's monitoring phase lasts 26 weeks, enrollment is expected to continue until approximately August 2025 and overall data collection should be concluded by the projected date of February 2026.

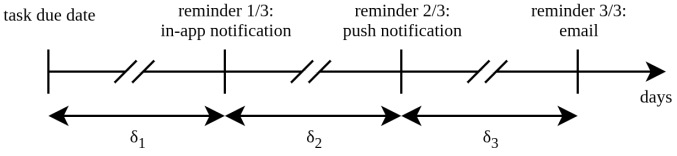
Engagement and Adherence Strategy

Sustaining participation in long-term decentralized trials presents a two-fold challenge: ensuring not only high participant adherence (the consistent use of devices and completion of surveys), but also deep engagement (the commitment to providing high-quality data). While traditional studies foster this commitment through personal contact, digital trials must employ different strategies[49,50] Our study employs a three-pronged digital strategy to address both. To cultivate the

engagement necessary for thoughtful survey responses and proper device use, we educate participants on the importance of data integrity and apply gamification techniques. In addition, adherence is directly rewarded through financial incentives.

The initial participant onboarding process via the WMA is designed to mitigate the occurrence of low-quality data. To this end, the communications first establish the importance of valid research outcomes before detailing the two central requirements for participants: correct and regular device usage, and the timely and accurate completion of questionnaires.

Additionally, scheduled nudges concerning the participant tasks (such as questionnaire completion, proper device usage and device synchronization) are disseminated. The nudges channel and timing have been designed to address the trade-off between securing timely, high-quality data and minimizing disruption to participants' daily lives. The schedule is available in Figure 2. When a task is detected as incomplete, a sequenced notification strategy is initiated, beginning with subtle in-app reminders and escalating to more direct channels like push notifications or emails if the task remains incomplete. More precisely, this sequence is triggered only after a set delay period of δ_1 days following the task's original due date and commences with an in-app notification within the WMA. If the task remains outstanding after a predetermined fixed temporal interval δ_2 days, a mobile device push notification is subsequently issued. Should the task still be incomplete after δ_3 days, an email communication is sent.



Task	δ_1	δ_2	δ_3
Use & sync of SW or SW2	7	10	30
Use & sync of WSA	2	7	30
Answer recurring survey	1	3	N/A
Answer survey I	1	6	N/A
Answer any other survey	1	6	N/A

Figure 2: Schedule of sequence of nudges for incomplete participant tasks
All parameters presented in the accompanying table for this figure are expressed in days.
N/A: Abbreviation for “Not Applicable”, indicating that no such reminder is scheduled for the specified task.

As gamification demonstrated to enhance adherence in health studies[51], the notification content integrated the participant progression, using the following metrics: the percentage completion of the individual's participation duration, the cumulative number of questionnaire invitations sent to the participant, and the achievement of specific temporal milestones, such as reaching 50% or 75% completion of their total study involvement.

Since previous studies encourage the concomitant use of gamification and financial incentives to foster overall adherence[51], a compensation has been incorporated. Eligibility for this financial benefit requires participants to meet two conjunctive criteria assessed over the participant's entire 26-week involvement: firstly, at least one of the designated study devices must be utilized and successfully upload data on more than 60 % of the nights (and 60 % of the days for the watches), and secondly, more than 60 % of all requested questionnaires must be completed. Eligible participants will be offered a choice between two forms of compensation: a 100€ Withings discount coupon or a six-month subscription to the Withings+ service. The protocol employs a simple, all-or-nothing reward model: participants receive the full compensation if they meet predefined data submission thresholds, and nothing otherwise. This structure was deliberately

chosen over more complex systems because it offers two key advantages: it is simpler to implement for the research team, and its rules are unambiguous for participants[52].

Withdrawal

To alleviate potential concerns regarding the long-lasting participation period and its substantial data collection demands, participants are informed of their unconditional right to withdraw consent at any time. The straightforward withdrawal process is detailed in the information notice and continuously reiterated as a reminder at the end of each questionnaire.

To withdraw, participants contact Withings Customer Support, via the designated email address or online Help Center, promptly receiving acknowledgement of their request's receipt and initiation of processing. The maximum withdrawal processing period of seven business days is outlined in the information notice and reiterated upon request.

For an effective execution compliant with participant wishes and General Data Protection Regulation (GDPR), participant withdrawals trigger a coordinated process. An automated participant identification system is utilized to prioritize communications from study participants over general customer inquiries. Customer Support immediately relays requests to the Data Protection Officer (DPO) and the Clinical Study Manager. The DPO ensures timely compliant handling of the request, while the Clinical Study Manager instructs the Data Management team to freeze or delete data as per the participant's request and halt all study communications using the CRM database. The Clinical Study Manager also provides final oversight, confirming all tasks are complete for formal closure.

Data Management

Given the scale and complexity of such a digital study, a robust data management plan is fundamental to yielding a high-fidelity dataset. This plan is designed to ensure the collection of high-quality and reliable data, while also upholding stringent standards for data integrity, security, privacy, and availability across the entire data lifecycle: from acquisition to final dataset preparation. This section outlines the comprehensive and fully automated procedures established to proactively address potential sources of data degradation or loss.

Data Security and Regulatory Compliance

To ensure rigorous data protection in this study, all data undergo pseudonymization and adheres to data minimization principles. Data storage occurs within secure, ISO 27001 certified environments, employing robust encryption and strict access controls, with all activities logged and monitored. Regular security audits and risk assessments are conducted to verify ongoing compliance. All data handling practices fully comply with the GDPR and applicable French data protection laws. Furthermore, Withings maintains ISO 27701 certification for its Privacy Information Management System.

Data Acquisition and Quality Assurance

Following informed consent, specific device functionalities identified as insightful for sleep assessment are remotely activated; these include, for the WSA, features enabling the detection of snoring and sleep apnea, and for the SW and SW2, the measurement of nocturnal oxygen

saturation (SpO2). However, recognizing that these advanced features can cause inconvenience, such as increased battery consumption or concerns regarding microphone activation, the study prioritizes long-term participant retention over obtaining the richest possible dataset from every individual. Participants are therefore explicitly informed of their right to disable these specific functions, a choice that provides a less granular dataset but ensures their continued participation with core data collection intact.

The management of device firmware (FW) versions is a critical aspect of maintaining data quality throughout the study. This involves continuously ensuring FW compatibility and consistency for all participants over the entire study duration. A key concern is mitigating potential risks to data uniformity that may arise from subsequent mandatory device updates, as these updates could alter the outputs of embedded data processing algorithms, thereby affecting the collected objective data. Instead of forcing participants onto a single, static FW version that would compromise user security and experience, our primary mitigation strategy is to meticulously record the firmware version associated with every data point. This logged version information is essential for careful subsequent analysis, allowing for stratification or adjustment based on any firmware variations. This is supplemented by strategically timing the launch of new study cohorts relative to anticipated software update schedules.

To mitigate risks to data quality, a screening questionnaire is administered before final enrollment. This questionnaire is designed to exclude individuals when there is a high likelihood of generating unreliable data, a risk identified from two potential sources: 1) problematic anticipated usage patterns, such as device sharing or Bluetooth misconfiguration, and 2) device malfunctions, identified using criteria derived from historical customer support data.

Because the recurring bi-weekly questionnaires assess the preceding 2-week period (Figures 1 and 2), prompt submission is essential to minimize recall bias. Therefore, participants are given a strict six-day response window to ensure their answers accurately reflect this recent timeframe. These windows also establish temporal consistency by standardizing data collection at regular, predefined intervals, which is critical for robust longitudinal analysis. This structured approach also minimizes participant confusion by preventing multiple, potentially conflicting, versions of a questionnaire within a brief timeframe. Questionnaires not completed within this six-day timeframe are recorded as missed. This system also supports the criteria for equitably rewarding timely and consistent data provision.

To address potential data loss from device connectivity interruptions (e.g., Bluetooth between SW/SW2 devices and smartphones, or Wi-Fi for WSA), which may still occur despite the participant screening measures previously described, an automated monitoring system detects extended periods of absent data transmission. Upon such detection, scheduled reminders are dispatched to participants. The frequency of these reminders is calibrated based on each device's internal storage capacity, aiming to prompt data synchronization before older, unsynced data is potentially overwritten by more recent recordings. These notifications also serve as general device usage reminders if recent activity is lacking.

To ensure critical data transmission integrity within this study's decentralized architecture, the automated data upload system's reliability is continuously verified using dedicated monitoring

tools. Transmission quality metrics have been established and tracked, including the volume of consecutive data points received, total data volume per participant, and the frequency and magnitude of data loss incidents.

Data Organization and Dataset Preparation

Previous studies[49] highlight the challenges associated with managing diverse data sources, such as questionnaire responses, device data, and withdrawal information. These challenges were addressed by implementing robust automated procedures. A primary function of these procedures is to ensure all collected data are accurately linked to a unique, pseudonymized participant identifier. These procedures also establish unambiguous temporal ordering for every data point, irrespective of its source.

To ensure timely data availability, minimize data loss risk, and facilitate ongoing quality control, the study dataset is constructed progressively throughout the research period. This iterative method necessitates meticulous upfront planning, including a detailed inventory of all data variables, their sources, retrieval mechanisms, and the defined dataset architecture. Every three days, an automated pipeline fetches new data from the user database, structures it according to this predefined architecture, and commits the output to a dedicated repository management by the open-source Git Large File Storage (LFS) extension[53] that serves as the study database. This process ensures robust version control and secure, incremental data aggregation.

Study Monitoring

Ongoing study monitoring is vital for sustaining research integrity, data quality, and participant engagement, as well as for early pipeline and dashboard debugging. A primary aspect of this process is the active management of recruitment bias. While the study's sampling population (the Withings users) is inherently distinct from the general population, thus creating a known selection bias, a key objective of our monitoring is to minimize an additional participation bias introduced by the recruitment protocol itself. To this end, we continuously benchmark the enrolled cohort's characteristics — including demographic, clinical, and device-usage data — against the broader user database to detect imbalances, enable timely investigation of protocol-related barriers and adjust recruitment strategies, ensuring the sample is representative of its source. This focus on the cohort's internal validity will be complemented by a final characterization of its composition and biases (e.g., insomnia prevalence), a critical prerequisite for valid subgroup comparisons and the development of robust machine learning models. This separate and broader challenge of generalizability to the general population will be addressed during the analysis phase.

Monitoring also guides other proactive and corrective measures, such as recruiting targeted underrepresented subgroups, addressing high rates of participants disabling some device functionalities, and proactively managing participant adherence through trend analysis and revision of motivational content. Daily data updates are first structured by a model running on the open-source Data Book Tool[54]. The resulting structured data is then fed into automatically updated dashboards on Metabase[55], another open-source platform. This data pipeline facilitates both data aggregation and visualization, to enable data tracking every second day. This monitoring encompasses a variety of metrics, systematically tracked across distinct domains of study execution and participant engagement.

a. **Recruitment:** Monitoring this aspect involves the following key assessments:
    - Number of enrollments against predefined estimations (as detailed in the Sample Size section), shown in Figure 3.
    - Estimated number of the remaining potential participants.
    - Proportion of the participants who have been excluded by the study's novel screening criteria. These criteria, developed from user issues reported via customer support, are specifically monitored to ensure they are not unduly restrictive.
    - Yield from different recruitment communication channels.
b. **Participant Cohort Characteristics:** To identify and address potential cohort imbalances, ongoing monitoring assesses the distribution of enrolled participants across demographic and clinical profiles, as well as two non-mutually exclusive device-usage groups — specifically, individuals utilizing at least one smartwatch (ScanWatch or ScanWatch 2) and those utilizing at least one Withings Sleep Analyzer. These distributions are then benchmarked against a predefined target population profile derived from the Withings CRM database. Illustrations are available in Figures 4 and 5.
c. **Questionnaire Engagement:** Questionnaire completion rates and intra-questionnaire drop-off points (see example in Figure 6).
d. **Adherence to Device Functionalities:** Rates at which participants maintain active use of, or choose to disable, the device functionalities automatically enabled at consent, as explained in *Data Acquisition and Quality Assurance* section.
e. **Compensation Eligibility:** Number of participants achieving compensation eligibility.

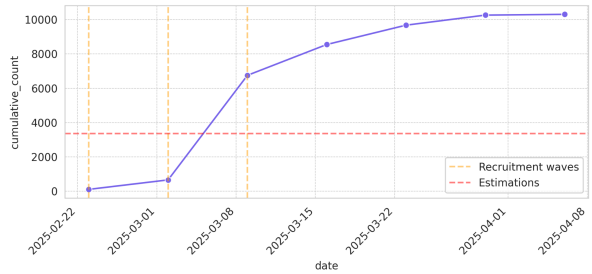


Figure 3: Comparison of actual and projected participant inclusions
Cumulative participant enrollment (solid blue line) is shown against the initial projection of ~3,400 (dashed red line).
The increases correspond to three recruitment waves targeting cohorts of approximately 2k, 10k, and 50k users, as detailed in the *Methodological Validation and Phased Implementation* section.

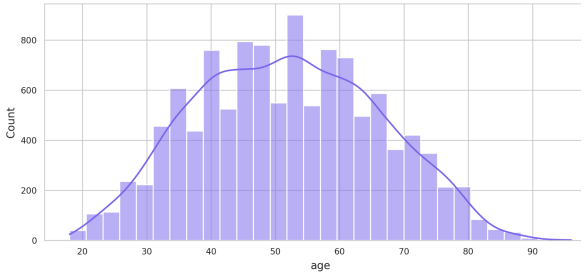


Figure 4: Distribution of participants age

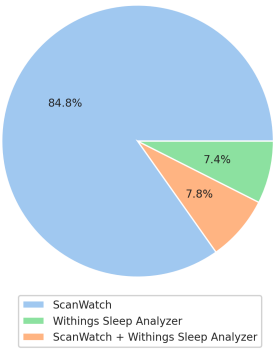


Figure 5: Distribution of participants owned devices

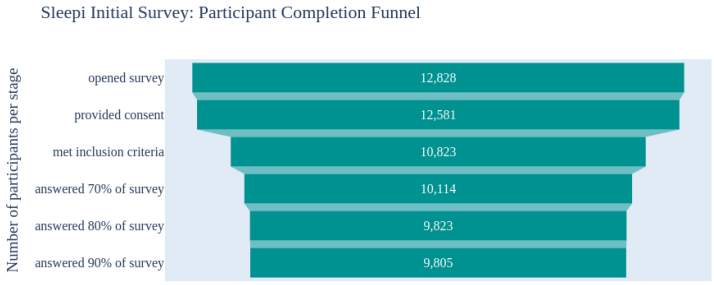


Figure 6: Illustration of participants drop-off points
A total of 62,000 users were contacted.

Sample size estimation: Determining necessary and achievable cohort size

The primary aim of this exploratory study is to generate a high-volume, longitudinal dataset of sufficient diversity to train ML models for insomnia assessment from the ground up. Because the objective is the discovery of robust, generalizable patterns rather than the testing of a single, pre-specified hypothesis, conventional sample size calculations are methodologically inappropriate for this data-driven research paradigm.

Determining the necessary sample size

Determining the necessary sample size to train an ML model is driven by the principles of statistical learning. A training dataset must be of sufficient scale and diversity, exposing the model to a wide variety of examples and configurations across all relevant subgroups, to capture the complexity of the problem, and therefore ensure it can develop robust and generalizable capabilities. To ensure an ML model can achieve high performance, we first need to estimate the required dataset size. To establish this estimate, we employed a composite approach grounded in theory, literature, and direct simulation.

While theory provides some guidance for simple classifiers[56,57], there is no universal rule for complex, non-linear models. The literature suggests that simpler models (e.g., SVM, Random Forests) often require about thousands of samples, while more complex architectures (e.g., XGBoost, Neural Networks) require tens of thousands[58–61]. These figures are rules of thumb, as the exact number is highly contingent on the specific properties of the dataset and the chosen model. Key dataset characteristics include the signal-to-noise ratio (SNR), overall dimensionality, the quantity and quality of features (including their correlation with the outcome), and the degree of class separability. This is further influenced by the specific parameters of the model architecture, which are themselves tuned based on the collected data.

To derive an estimate specific to our use case, we conducted an a priori power analysis using Monte Carlo simulations[62,63]. This standard approach involves a sequence of steps: first, we defined a data-generating process to create synthetic training and testing sets that mimic our expected data. Second, a success criterion was established, defined as the ability of a model to distinguish between insomnia and healthy subjects with a given Area Under the Curve (AUC) on the test set.

This process was then simulated repeatedly for a given sample size. This allows for the calculation of the statistical power, which in this context is defined as the proportion of simulation runs where the model achieved its success criterion. Finally, this entire procedure was iterated across different sample sizes to identify the minimum required to achieve our target power. The detailed parameters for the data generation process, models, and success criteria are available in the Supplementary Materials B.

The simulation results (Figure 7) indicate the sample sizes required to achieve a statistical power of at least 80% for different performance targets. For a target AUC ≥ 0.94, an SVM requires approximately 800 participants and a Neural Network requires 1,100. For a target of AUC ≥ 0.90, an XGBoost model requires 800 participants, while a Random Forest model needs approximately 3,200. The detailed results for sensitivity and specificity at these sample sizes are available in the Supplementary Materials B.

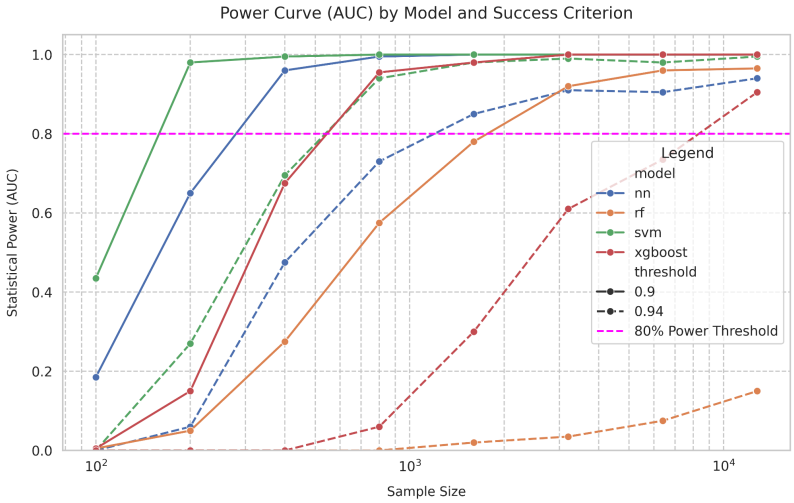


Figure 7: Power curves obtained through simulation for insomnia classification using different AUC success criteria. nn: neural network, rf: random forest, svm: support vector machine

Based on these findings, and supported by theory and literature, we concluded that a sample size of about 1,000 participants is sufficient to reasonably train a standard machine learning model, such as an SVM, and achieve satisfying performance for this task.

Estimating the achievable sample size: a prospective attrition model

Having established a target for the necessary sample size, the next critical step was to determine if this target was practically achievable and to calculate the required number of initial contacts. To do this, we developed a prospective attrition model, a well-established approach in clinical research for forecasting participant loss.

This model provides a practical, ex-ante estimation of the final sample size by applying conservative retention rates at each sequential stage of the study: recruitment, enrollment, and the 26-week active participation period. These retention rates, defined as the proportion of participants expected to progress to the subsequent stage, are reportedly derived with a conservative approach. The derivation draws upon two primary sources: empirical data from large-scale internal Withings studies, which involved user populations ranging from 10,000 to

over 1,000,000 subjects, and statistics derived from general user engagement observed in the broader Withings user base. We accounted for all identified causes of data loss in these estimations.

The attrition stages, their definitions, the derivation of the data loss risk assessments, their corresponding retention rates, the detailed sequential calculation methodology are available in Supplementary Materials A. After sequentially applying these estimated retention rates through the entire recruitment and participation pathway, the model projects that about 83,662 users need to be contacted to include 3,175 participants in the study and eventually yield 1,000 usable data for the analyses (Figure 8).

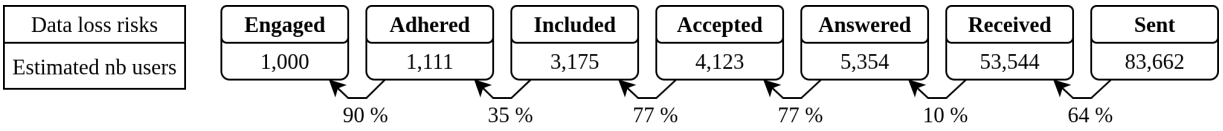


Figure 8: Illustration of the prospective attrition model

Conclusion

This estimation process provides a robust confirmation of the study's feasibility. The initial Monte Carlo power analysis established that a target sample size of approximately 1,000 participants is necessary to complete our machine learning objectives. Subsequently, the prospective attrition model confirmed that this target is practically achievable, projecting that an initial outreach to approximately 84,000 users would be required to yield the final cohort. Crucially, this required number of contacts is within the available sampling frame of active Withings users in France, validating that the study design is appropriately scaled to meet its data collection goals.

Methodological Validation and Phased Implementation

To ensure the robustness of study procedures prior to and during full-scale operation, a sequence of validation steps and a phased implementation approach were adopted. The different phases are summarized in Table 1.

Testing Phases

The value of pilot testing in large-scale digital health research, particularly for validating data systems including data flow, integration, and integrity, is well-established[49]. Informed by such findings and to meticulously prepare for the main trial, two distinct phases of testing and protocol refinement were conducted exclusively with Withings employees.

The first phase of technical validation involved 18 participants for one month. Its objectives were to identify and resolve significant bugs in automated study processes. Testing encompassed the questionnaire and reminder deployment system, detailing its scheduling and conditional logic; the data management pipeline, specifying its functions for data acquisition, integrity checks, linkage, and initial dataset building routines; and device feature activation protocols. This initial phase also finalized the development and initial validation of the study monitoring dashboards.

Subsequently, a larger refinement phase was conducted with 80 participants, also for one month. An essential characteristic of this cohort was that the majority of individuals had no prior knowledge of the study's specific existence, goals, or design, making them ideally suited for this phase's primary focus: assessing and refining the clarity of all participant-facing materials. Particular attention was given to ensuring that study criteria, device usage guidelines, and compensation information were unambiguous for individuals new to the study's specific requirements. Further objectives included generating preliminary attrition estimates, evaluating initial engagement with the long baseline questionnaire, and assessing early adherence patterns and associated questionnaire burden from the first recurring questionnaires.

The collective findings from these validation and refinement testing phases directly informed the finalization of the study protocol and operational workflows before commencing large-scale participant recruitment.

Initial Study Rollout and Operational Validation

Following testing phases, large-scale recruitment was initiated, rolling out to 12,000 targeted individuals in two successive waves: an initial 2,000, followed by 10,000. While this phase marked the formal start of data collection for the primary study objectives, it concurrently served as a critical period for real-world operational validation and refinement at scale.

The objectives during this initial scaled implementation were to:

1. Verify the accuracy of participant targeting achieved through the internal CRM tool's population segmentation.
2. Assess the performance, stability, and scalability of all study systems under increased load, including recruitment pipelines, participant engagement throughout the initial long questionnaire, and adherence to recurring questionnaire schedules.
3. Further refine the participant withdrawal process based on real-world usage and feedback.
4. Identify and promptly resolve any residual technical issues or bugs specific to the production environment before extending recruitment to the remainder of the target population.

This phased approach allowed for iterative improvements and risk mitigation during the early stages of the main study.

Phase Name	Sampling Population	# Targeted Subjects	Time to respond	Objectives
Technical Validation	Withings employees	18	1 month	• Identify and resolve bugs in automated systems (questionnaires, data pipeline). • Validate device feature activation. • Finalize study monitoring dashboards.
Protocol Refinement	Withings employees	80	1 month	• Assess and refine the clarity of all participant-facing materials. • Generate preliminary attrition estimates. • Evaluate initial engagement, adherence, and questionnaire burden.
Initial Study Rollout & Operational Validation	Withings users	2,000	1 week	• Verify the accuracy of CRM-based participant targeting at scale. • Assess system performance, stability, and scalability under real-world load. • Refine the withdrawal process based on user feedback.
	Withings users	10,000	1 week	
Main Study Launch	Withings users	50,000	6 weeks	• Execute the main phase of long-term data collection with the initial device cohorts (SW and WSA).
Final Study Launch	Withings users	25,000	N/A	• Execute the main phase of long-term data collection with remaining device cohort (SW2).

Table 1: Summary of Study Validation and Phased Implementation

The study's recruitment is deliberately structured in distinct phases to manage firmware updates across devices. For instance, a group of 25,000 SW2 users is reached out to in a later recruitment wave because of the necessity to release a new device firmware and to give enough time for the users to make the update.

Study Organization

Withings sponsors this clinical trial, providing funding and operational infrastructure. Guided by diverse expertise and a structured organization, the study's execution upholds rigorous conduct, data integrity, participant protection, and regulatory compliance. The Principal Investigator is an external clinician and sleep research specialist from leading French public institutions including Inserm, Université Paris Cité, and the Paris Public Hospital Network (AP-HP, Hôpital Bichat). All contributing parties share responsibility for the appropriate design, conduct, analysis, and interpretation of this study's findings.

Enrollment Status

The SLEEPI study was initiated on February 18, 2025. During the initial phase of participant recruitment, 62,000 users within the Withings ecosystem were contacted, resulting in 10,823 unique individuals included within the first six weeks. This represents a contact-to-inclusion conversion rate of approximately 17%, with only four participants having withdrawn.

These inclusion figures significantly surpassed the study's original projections. This initial projection, formulated at the time of protocol design, had anticipated a total of 3,354 inclusions

over a 26-week period. Individuals who owned multiple devices were counted more than once. The projection was based on an estimated total pool of approximately 84,000 contacted users, which included the ScanWatch 2 cohort, who were excluded from the initial recruitment wave. For the detailed calculations about the projections, please refer to Supplementary Materials A.

This higher-than-projected enrollment (Figure 3) stemmed from three key stages where the initial attrition model proved overly conservative:

- Initial survey response: The response rate to the initial outreach was significantly higher than anticipated (16% vs. a projected 10%), likely driven by strong user interest in sleep research and the provided financial incentive.
- Consent rate: Among those who responded, the rate of consent to participate was exceptionally greater than expected (96% vs. a projected 77%).
- Inclusion rate: Following consent, the rate of individuals meeting the final inclusion criteria was also higher than expected (90% vs. a projected 81%).

The protocol's high efficiency was demonstrated during the initial recruitment phase, which significantly surpassed projections and prompted a temporary pause to revise the study budget. This success validated the framework's scalability. The system has proven capable of absorbing a much larger participant influx than planned, confirming that the primary constraint is budgetary, not operational.

Discussion

Limitations

Challenges of Datasets Derived from Consumer Devices

The use of consumer-grade devices within free-living conditions is a deliberate methodological choice that prioritizes scalability, user experience, cost-effectiveness, and longitudinal data collection over the precision attained in a clinically controlled environment. To enable passive, automatic, long-term collection of data, consumer smart devices sometimes use proxies (e.g. pressure variations in a bladder caused by chest movements in place of impedance belts) and a less comprehensive set of sensors than some gold standard techniques like the PSG. Furthermore, variable participant adherence to device use instructions leads to more noise and missing values in decentralized trials than in centralized studies. Consequently, occurrences of data loss or intermittent gaps are anticipated, necessitating significant preprocessing, cleaning, and potentially imputation efforts to prepare a robust dataset suitable for analysis.

Furthermore, the study cohort, composed primarily of existing Withings device users, may not fully mirror the broader general population. This sampling approach can introduce selection biases; for example, the cohort might differ in terms of socio-economic status, technological proficiency, or baseline health awareness. Such biases could limit the generalizability of observed patterns and outcomes to more diverse demographic groups.

Trial Digitalization: A Foundational, Resource-Intensive Infrastructure

Maintaining long-term participant adherence and engagement in this decentralized study presents significant challenges for personalized, proactive intervention, in contrast to more traditional site-based study designs that rely on direct contact with healthcare professionals for support and assessment. Consequently, the study relied heavily on exceptionally clear study materials, precise instructions, and carefully designed automated reminders to guide participants, while strategically balancing information provision to avoid overwhelming them.

Implementing this decentralized study necessitated substantial initial investment to develop and adapt existing Withings tools and resources into reusable foundational infrastructure that includes robust data management systems, adaptable study design protocols, and defined workflows covering recruitment, participant withdrawal, and phased implementation strategies. Although the initial setup was resource-intensive, these core developments offer significant future reusability across various health domains.

Future Works

The methodology for decentralized trials could be significantly advanced through subsequent research in several areas identified by this study. Firstly, developing an in-app participant interface for real-time progress tracking could improve clarity and support more advanced gamification strategies. Secondly, systematic A/B testing is recommended to empirically optimize engagement approaches, such as CRM content and reminder protocols[64,65]. Thirdly, the rich longitudinal dataset collected warrants in-depth analysis to characterize long-term patterns of participant adherence and engagement with both devices and questionnaires, thereby identifying predictors of sustained participation. Finally, future protocols could incorporate methods to characterize and statistically adjust for the selection biases inherent to commercial user populations, thereby enhancing the generalizability of findings to the broader public.

Conclusion

This study presents a methodological framework for a large-scale, decentralized trial using consumer smart devices to identify digital biomarkers for subjective insomnia. This protocol addresses the increasing need for large, longitudinal datasets essential for advanced analytics like machine learning approaches, a requirement often challenging for traditional research paradigms and sample size methodologies.

Methodological contributions include a two-part sample size strategy that first uses simulation to define the minimum sample required for the machine learning objective, and then employs a prospective attrition model, based on clinical trial forecasting[66], to calculate the number of initial contacts needed to achieve that target. Furthermore, the overall design for this large-scale, digital, longitudinal and decentralized trial offers a replicable and scalable template covering critical aspects such as recruitment from an existing user base, data quality assurance, participant engagement, and robust data management.

While ambitious decentralized studies face inherent challenges detailed in the Limitations section, this work provides practical insights and transferable principles for proactive and efficient study management and monitoring. While this trial benefited from an established digital ecosystem, its core methodologies are adaptable for other organizations conducting large-scale real-world data studies. However, such adaptation would necessitate a careful assessment of the distinct resource requirements — particularly for participant recruitment, device provisioning, and ongoing support — especially for entities lacking a comparable pre-existing infrastructure.

Ultimately, this methodological development is driven by the clinical need for more nuanced, objective tools to support specialists in insomnia diagnosis and care. By seeking digital biomarkers grounded in data that can be correlated with clinically validated subjective assessments[7,45–47], the aim is to develop technologies that complement existing clinical pathways. Thus, this study not only contributes to insomnia research but also demonstrates a scalable methodological approach for translating real-world digital health data into clinically relevant insights, with potential applicability to other health domains.

Data Availability

The data collected in this study were collected during a clinical study (Approval Reference Number: 2024-A01028-39, French Institutional Review Board) conducted by Withings. The dataset associated with this study is stored in a proprietary repository (Withings) and cannot be shared publicly due to concern for privacy, ethical and legal reasons. Queries for data access can be directed to Withings (privacy@withings.com) with a timeframe for response of four weeks.

Funding Declaration

This study is funded by Withings.

Acronyms

BMI: Body mass index
CPP: Comités de Protection des Personnes (French Ethics Committee)
CRM: Customer Relationship Management
DHTs: Digital Health Technologies
DSM-5-TR: *Diagnostic and Statistical Manual of Mental Disorders, Text Revision*
GAD-7: Generalized Anxiety Disorder-7
GDPR: General Data Protection Regulation
HRV: Heart Rate Variability
ISI: Insomnia Severity Index (questionnaire)
LFS: Large File Storage
ML: Machine Learning
PHQ-9: Patient Health Questionnaire-9

PRO: Patient-Reported Outcomes
PSG: Polysomnography
SLEEPI: Smart devices-based Longitudinal Assessment of Patterns of Insomnia (clinical trial full title)
WMA: Withings Mobile Application
WSA: Withings Sleep Analyzer
SW(2): ScanWatch (2)

REFERENCES

1. Khozin, S. & Coravos, A. Decentralized Trials in the Age of Real-World Evidence and Inclusivity in Clinical Investigations. *Clin. Pharmacol. Ther.* **106**, 25–27 (2019).
2. Decentralized Clinical Trials for Drugs, Biological Products, and Devices; Draft Guidance for Industry, Investigators, and Other Stakeholders; Availability. *Federal Register* https://www.federalregister.gov/documents/2023/05/03/2023-09399/decentralized-clinical-trials-for-drugs-biological-products-and-devices-draft-guidance-for-industry (2023).
3. Turakhia, M. P. *et al.* Rationale and design of a large-scale, app-based study to identify cardiac arrhythmias using a smartwatch: The Apple Heart Study. *Am. Heart J.* **207**, 66–75 (2019).
4. Detection of Atrial Fibrillation in a Large Population Using Wearable Devices: The Fitbit Heart Study | Circulation. https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.122.060291.
5. Perez, M. V. *et al.* Large-Scale Assessment of a Smartwatch to Identify Atrial Fibrillation. *N. Engl. J. Med.* **381**, 1909–1917 (2019).
6. Riemann, D. & others. The European Insomnia Guideline: An update on the diagnosis and treatment of insomnia 2023. *J. Sleep Res.* **32**, e14035 (2023).
7. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (5th Ed., Text Rev.)*. (2022).

8. Leger, D. & others. Prevalence of insomnia in a survey of 12,778 adults in France. *J. Sleep Res.* **9**, 35–42 (2000).
9. Okajima, I. *et al.* Evaluation of Severity Levels of the Athens Insomnia Scale Based on the Criterion of Insomnia Severity Index. *Int. J. Environ. Res. Public. Health* **17**, 8789 (2020).
10. Morin, C. M. & Jarrin, D. C. Epidemiology of Insomnia: Prevalence, Course, Risk Factors, and Public Health Burden. *Sleep Med. Clin.* **17**, 173–191 (2022).
11. Leger, D. & Poursain, B. An international survey of insomnia: under-recognition and under-treatment of a polysymptomatic condition. *Curr. Med. Res. Opin.* **21**, 1785 (2005).
12. Biedebach, L. & others. Two Sides of the Same Pillow: Unfolding the Relationship between Objective and Subjective Sleep Quality with Unsupervised Learning. (2023).
13. Castro, L. S. & others. Objective prevalence of insomnia in the São Paulo, Brazil epidemiologic sleep study. *Ann. Neurol.* **74**, 537–546 (2013).
14. Reed, D. L. & Sacco, W. P. Measuring sleep efficiency: what should the denominator be? *J. Clin. Sleep Med.* **12**, 263–266 (2016).
15. Geoffroy, P. A. *et al.* Sleep in remitted bipolar disorder: A naturalistic case-control study using actigraphy. *J. Affect. Disord.* **158**, 1–7 (2014).
16. Romier, A. *et al.* Objective sleep markers of suicidal behaviors in patients with psychiatric disorders: A systematic review and meta-analysis. *Sleep Med. Rev.* **68**, 101760 (2023).
17. Rundo, J. V. & Downey, R. I. Polysomnography. *Handb. Clin. Neurol.* **160**, 381–392 (2019).
18. Kang, S.-G. & others. Validity of a commercial wearable sleep tracker in adult insomnia disorder patients and good sleepers. *J. Psychosom. Res.* **97**, 38–44 (2017).

22

19. Cheng, Y.-H., Lech, M. & Wilkinson, R. H. Simultaneous Sleep Stage and Sleep Disorder Detection from Multimodal Sensors Using Deep Learning. *Sensors* **23**, (2023).

20. Hamida, S. T.-B., Ahmed, B. & Penzel, T. A Novel Insomnia Identification Method Based on Hjorth Parameters. in *2015 IEEE INTERNATIONAL SYMPOSIUM ON SIGNAL PROCESSING AND INFORMATION TECHNOLOGY (ISSPIT)* 548–552 (2015).

21. Zhou, L. & others. Machine learning on big data: Opportunities and challenges. *Neurocomputing* **237**, 350–361 (2017).

22. Vallières, A. & Morin, C. M. Actigraphy in the Assessment of Insomnia. *Sleep* **26**, 902–906 (2003).

23. Natale, V., Plazzi, G. & Martoni, M. Actigraphy in the Assessment of Insomnia: A Quantitative Approach. *Sleep* **32**, 767–771 (2009).

24. Natale, V., Léger, D., Martoni, M., Bayon, V. & Erbacci, A. The role of actigraphy in the assessment of primary insomnia: a retrospective study. *Sleep Med.* **15**, 111–115 (2014).

25. Lichstein, K. L. *et al.* Actigraphy validation with insomnia. *Sleep* **29**, 232–239 (2006).

26. Afshani, M. & others. Discriminating paradoxical and psychophysiological insomnia based on structural and functional brain images: a preliminary machine learning study. *Brain Sci.* **13**, 672 (2023).

27. Kusmakar, S. & others. A machine learning model for multi-night actigraphic detection of chronic insomnia: development and validation of a pre-screening tool. *R. Soc. Open Sci.* **8**, 202264 (2021).

28. Óskarsdóttir, M. & others. Importance of getting enough sleep and daily activity data to assess variability: longitudinal observational study. *JMIR Form. Res.* **6**, e31807 (2022).

29. Bitkina, O. V., Park, J. & Kim, J. Modeling Sleep Quality Depending on Objective

Actigraphic Indicators Based on Machine Learning Methods. *Int. J. Environ. Res. Public. Health* **19**, (2022).

30. Shahin, M. & others. Deep learning and insomnia: assisting clinicians with their diagnosis. *IEEE J. Biomed. Health Inform.* **21**, 1546–1553 (2017).

31. Sharma, M., Dhiman, H. S. & Acharya, U. R. Automatic identification of insomnia using optimal antisymmetric biorthogonal wavelet filter bank with ECG signals. *Comput. Biol. Med.* **131**, 104246 (2021).

32. Mayfield, J. J. *et al.* Implementation of a fully remote randomized clinical trial with cardiac monitoring. *Commun. Med.* **1**, 62 (2021).

33. Center for Drug Evaluation and Research. Digital Health Technologies for Remote Data Acquisition in Clinical Investigation.

34. Spina, M.-A., Andrillon, T., Wiley, J. F., Rajaratnam, S. M. W. & Bei, B. Does changing perceptions of sleep by incorporating sleep wearables improve insomnia? Protocol for a randomized study (the Novel Insomnia Treatment Experiment). *SLEEP Adv.* **4**, zpad012 (2023).

35. Stern, E. & others. How Can Digital Mental Health Enhance Psychiatry? *The Neuroscientist* 10738584221098603 (2022).

36. Yin, J., Xu, J. & Ren, T.-L. Recent Progress in Long-Term Sleep Monitoring Technology. *Biosensors* **13**, 395 (2023).

37. Lechat, B. & others. Multinight prevalence, variability, and diagnostic misclassification of obstructive sleep apnea. *Am. J. Respir. Crit. Care Med.* **205**, 563–569 (2022).

38. Edouard, P. & others. Validation of the Withings Sleep Analyzer, an under-the-mattress device for the detection of moderate-severe sleep apnea syndrome. *J. Clin. Sleep Med.* **17**,

1217–1227 (2021).

39. Jagielski, J. T. *et al.* Evaluating an under-mattress sleep monitor compared to a peripheral arterial tonometry home sleep apnea test device in the diagnosis of obstructive sleep apnea. *Sleep Breath.* **27**, 1433–1441 (2023).

40. Huyett, P. *et al.* Comparison of Home Sleep Devices and Sleep Study Testing in Hypoglossal Nerve Stimulation Patients. *The Laryngoscope* **135**, 469–477 (2025).

41. G Ravindran, K. K. *et al.* Reliable Contactless Monitoring of Heart Rate, Breathing Rate, and Breathing Disturbance During Sleep in Aging: Digital Health Technology Evaluation Study. *JMIR MHealth UHealth* **12**, e53643 (2024).

42. Yang, R.-Y. & others. Pneumatic sensor for cardiorespiratory monitoring during sleep. *Biomed. Phys. Eng. Express* **5**, 055014 (2019).

43. Ravindran, K. K. & others. Contactless and longitudinal monitoring of nocturnal sleep and daytime naps in older men and women: a digital health technology evaluation study. *Sleep* **46**, zsad194 (2023).

44. Stefanos, M.-A. *et al.* Evaluation of a Contactless Sleep Monitoring Device for Sleep Stage Detection against Home Polysomnography in a Healthy Population. 2025.05.06.25326860 Preprint at https://doi.org/10.1101/2025.05.06.25326860 (2025).

45. Bastien, C. H., Vallières, A. & Morin, C. M. Validation of the Insomnia Severity Index as an outcome measure for insomnia research. *Sleep Med.* **2**, 297–307 (2001).

46. Patient Health Questionnaire-9. https://psycnet.apa.org/doiLanding?doi=10.1037%2Ft06165-000.

47. Spitzer, R. L., Kroenke, K., Williams, J. B. W. & Löwe, B. Generalized Anxiety Disorder 7. https://doi.org/10.1037/t02591-000 (2011).

25

48. Article L1121-6 - Code de la santé publique - Légifrance. https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000025457547/2025-07-16/.

49. Garcia, A. *et al.* Lessons learned in the Apple Heart Study and implications for the data management of future digital clinical trials. *J. Biopharm. Stat.* **32**, 496–510 (2022).

50. Robbins, R. *et al.* A clinical trial to evaluate the dayzz smartphone app on employee sleep, health, and productivity at a large US employer. *PloS One* **17**, e0260828 (2022).

51. Tran, S., Smith, L., El-Den, S. & Carter, S. The Use of Gamification and Incentives in Mobile Health Apps to Improve Medication Adherence: Scoping Review. *JMIR MHealth UHealth* **10**, e30671 (2022).

52. Whiteley, L., Brown, L. K., Mena, L., Craker, L. & Arnold, T. Enhancing health among youth living with HIV using an iPhone game. *AIDS Care* **30**, 21–33 (2018).

53. Git Large File Storage. *Git Large File Storage* https://git-lfs.com/.

54. dbt Labs | Transform Data in Your Warehouse. *dbt Labs* https://www.getdbt.com/index.

55. Open source business intelligence, dashboards, and data visualizations | Metabase. https://www.metabase.com/.

56. Vapnik, V. N. & Chervonenkis, A. Y. (Translated by B. Seckler).

57. Riley, R. D. *et al.* Calculating the sample size required for developing a clinical prediction model. *BMJ* m441 (2020) doi:10.1136/bmj.m441.

58. Sample Size Requirements for Popular Classification Algorithms in Tabular Clinical Data: Empirical Study - PMC. https://pmc-ncbi-nlm-nih-gov.proxy.insermbiblio.inist.fr/articles/PMC11688588/.

59. Lee, L.-H. *et al.* Evaluating the performance of machine learning models for automatic diagnosis of patients with schizophrenia based on a single site dataset of 440 participants. *Eur.*

Psychiatry **65**, e1 (2021).

60. Floares, A. G. *et al.* The smallest sample size for the desired diagnosis accuracy. *Int J Oncol Cancer Ther* **2**, 6 (2017).

61. van der Ploeg, T., Austin, P. C. & Steyerberg, E. W. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med. Res. Methodol.* **14**, 137 (2014).

62. Constantin, M. A., Schuurman, N. K. & Vermunt, J. K. A general Monte Carlo method for sample size analysis in the context of network models. *Psychol. Methods* https://doi.org/10.1037/met0000555 (2023) doi:10.1037/met0000555.

63. Zimmer, F. & Debelak, R. Simulation-based design optimization for statistical power: Utilizing machine learning. *Psychol. Methods* **30**, 513–536 (2025).

64. Austrian, J. *et al.* Applying A/B Testing to Clinical Decision Support: Rapid Randomized Controlled Trials. *J. Med. Internet Res.* **23**, e16651 (2021).

65. Lee, H. *et al.* The Impact of Intervention Design on User Engagement in Digital Therapeutics Research: Factorial Experiment With a Mixed Methods Study. *JMIR Form. Res.* **8**, e51225 (2024).

66. Bogin, V. Lasagna's law: A dish best served early. *Contemp. Clin. Trials Commun.* **26**, 100900 (2022).

4.3 Chapter Summary

- SLEEPI is a large-scale decentralized clinical-trial protocol whose primary objective is to train and evaluate a machine learning model that separates individuals with insomnia from healthy sleepers using multivariate sensor data collected longitudinally under sparse, irregular sampling, with a secondary objective of assessing participant adherence.
- Two paired data streams: objective physiological data from three Withings devices (the WSA nearable and the SW and SW 2 smartwatches, recording sleep architecture, activity, temperature, HR, HRV, oxygen saturation, and snoring) and subjective questionnaires delivered in the Withings app (DSM-5 insomnia criteria, ISI, PHQ-9, GAD-7, plus demographic, clinical, and environmental data).
- Ground truth and tasks: patient-reported outcomes supply the labels, with DSM-5 criteria defining the binary insomnia/healthy classification and the ISI defining a severity-regression task; the training data are physiological biomarkers computed from the raw signals rather than the raw acceleration or photoplethysmography traces.
- Participant journey: each participant is monitored for 26 weeks under free-living conditions using at least one device and completing scheduled questionnaires, with recruitment, engagement, adherence, monitoring, data management, and withdrawal automated and run with minimal staff.
- Sample sizing: a two-part strategy first uses simulation to define the minimal sample required for the machine learning objective, then a prospective attrition model to compute the number of initial contacts needed to reach it.
- Results: continuous monitoring of more than 10,000 individuals over 26 weeks, the largest and most diverse longitudinal insomnia dataset to date, with enrollment exceeding projections (survey response 16% vs 10%, consent 96% vs 77%, inclusion 90% vs 81%) and validating scalability.
- Using consumer-grade devices in free-living conditions is a deliberate choice that prioritizes scale and long-term collection over clinical precision; the devices rely on proxies and a smaller sensor set than PSG, and variable participant adherence introduces noise and missing values that call for substantial preprocessing and imputation before analysis.
- The cohort, drawn mainly from existing Withings users, may differ from the general population in characteristics such as technological proficiency and socio-economic status, which limits the generalizability of the findings.
- Sustaining engagement without site-based clinical contact relied on clear materials and automated reminders.
- Take-home: the protocol offers a replicable, scalable template for large-scale real-world data collection, spanning recruitment from an existing user base, data-quality assurance, participant engagement, and data management, and it is adaptable to other organizations and health domains subject to their resource and infrastructure constraints.

Chapter 5

Subjective and objective characterization of insomnia in the SLEEPI cohort

5.1 Data quantity

The initial survey drew 11,821 respondents before the eligibility gate. Figure 5.1 begins after that gate, taking the 11,663 respondents who met the eligibility criteria as its baseline rather than this initial total. Of these, 11,341, or 97%, gave informed consent, and completion then declined steadily across the successive sections to 9,825 at the closing ISI questionnaire.

Across the three screening gates, 937 respondents were screened out: 158 at the eligibility gate, comprising 92 minors or respondents without French social-security cover and 66 vulnerable persons; 322 who declined consent; and 457 who had neither a usable ScanWatch nor a usable Sleep Analyzer. The two most frequent device-screening cases were the absence of a ScanWatch and frequent missing ScanWatch data, each accounting for 174 of these exits. The detail is given in Appendix D.1.

The analyses in this chapter draw on the resulting analyzable cohort of 10,878 users: the 10,884 who passed all three screening gates, less the 6 who later withdrew their consent. These six are retained in the recruitment funnel above, since they had completed screening, but are excluded from the analyzable cohort and from Table 5.1.

The ISI questionnaire was re-administered every two weeks across the 26-week monitoring period. Against the 10,884 screened cohort, participation began at 90.3% at week 0, fell to a low of 45.6% at week 16, and recovered to 52.7% by week 26, as Figure 5.2 shows.

Table 5.1 summarizes the per-participant data volume. On average, participants contributed 152 days with sleep data and 140 days with step data over the 182-day monitoring period, and answered the recurring ISI questionnaire 8 times out of the 14 administered. Sleep data are collected by both the WSA and the SW, so every participant with device

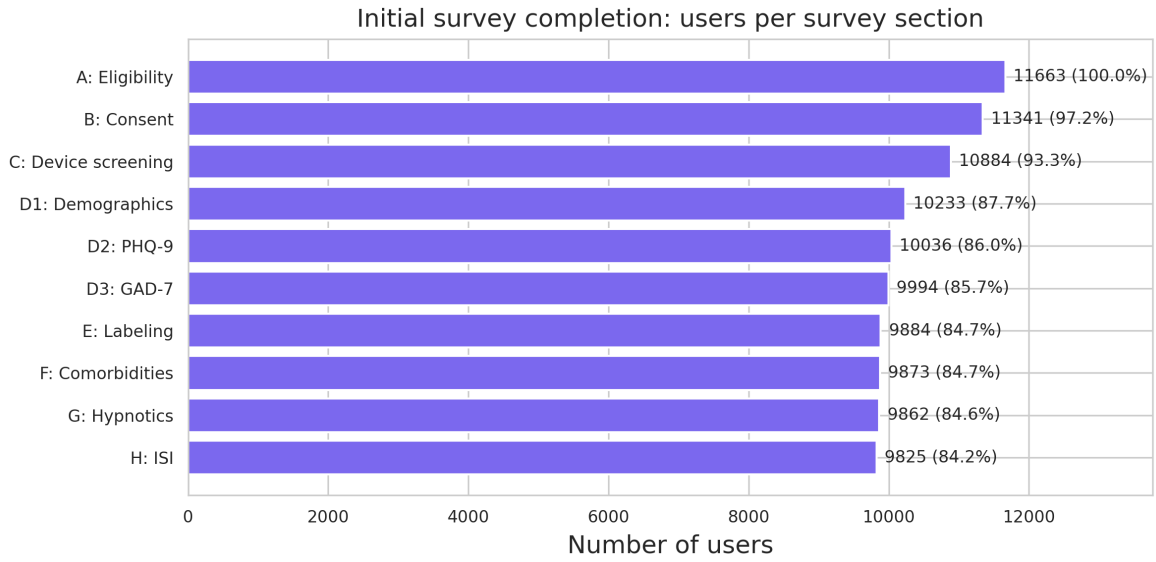


Figure 5.1: Initial-survey completion: the number of users reaching each section of the survey, from eligibility to the ISI questionnaire.

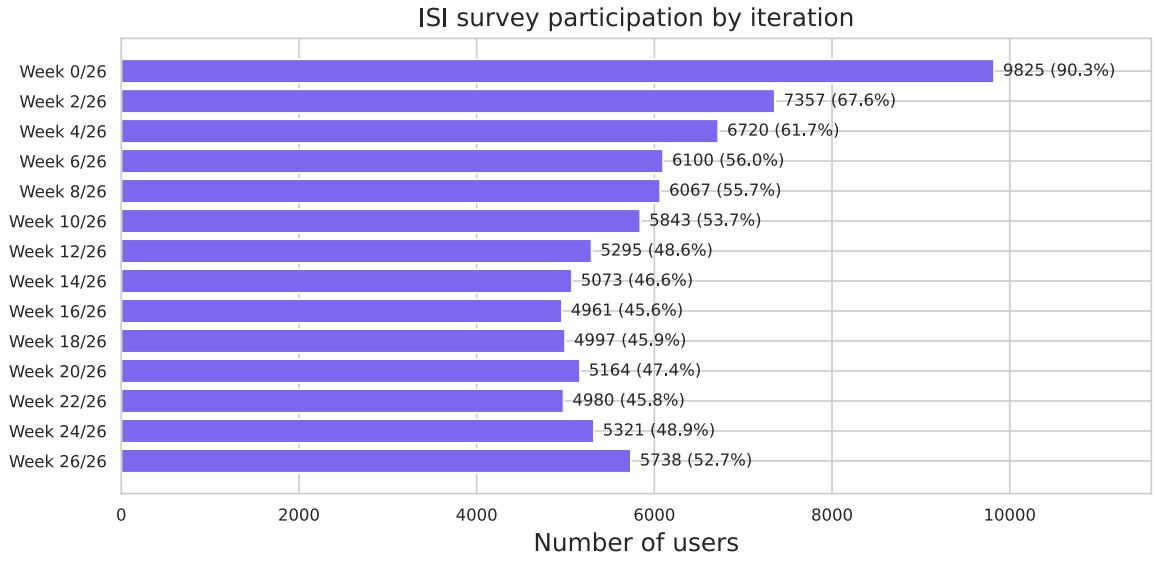


Figure 5.2: Participation in the recurring ISI questionnaire at each two-week iteration of the 26-week monitoring period.

recordings has nights of sleep data, whereas step data are collected only by the SW, so some participants have no step days at all, the zero shown as the 5th percentile of step-days in Table 5.1.

A participant qualified for the analysis dataset when three data-contribution criteria were met. First, at least one device was used over more than 60% of the study duration: for the SW and SW 2 wearables, on more than 60% of both the days and the nights, and for the WSA nearable, on more than 60% of the nights. Second, each section of the initial survey was completed at least once: the inclusion and exclusion criteria (A), the information letter

and informed-consent form (B), device-usage habits (C), demographics (D), labeling (E), comorbidities (F), and hypnotic use (G). Third, the ISI questionnaire (H) was answered at least seven of the fourteen times it was administered. Of the 10,674 users for whom these criteria could be computed, 5,478 met them, a 51% qualifying rate. Participants who met these criteria were eligible for the study compensation defined in the protocol described in Chapter 4.

Metric	n	p5	Mean	SD	p95
Days with sleep data	10,674	46	152	43	181
Days with step data	10,674	0	140	60	182
Recurring ISI answers	10,878	0	8	5	14

Table 5.1: Per-participant data-quantity metrics across the analyzable cohort. p5 and p95 are the 5th and 95th percentiles; SD is the standard deviation; n is the number of cohort users with the metric available. The device-day metrics are over the 10,674 participants with device recordings, whereas the recurring-ISI count is over the full cohort and is zero for participants who never answered. The maxima possible are 182 device-days over the 26-week monitoring period and 14 recurring-ISI administrations.

5.2 Demographics

The demographic and clinical analyses in this chapter use the analyzable cohort of 10,878 participants defined above.

Table 5.2 summarizes the continuous demographic variables, each over the participants who reported a plausible value. A survey bug detected and fixed during the first inclusion phases prevented some participants from entering a plausible height or weight, recording a year-like value instead; these 132 entries were readily identified and dropped as missing data. The mean height was 174 cm (SD 9), the mean weight 80 kg (SD 17), and the mean age 52 years (SD 14); the mean BMI, over the 8,121 participants with both height and weight, is 27 kg/m^2 (SD 5). Separately from these dropped entries, weight is reported by far fewer participants than height or age: roughly 1,900 respondents who gave their height and birth year have no recorded weight, the result of an upstream capture failure rather than participant skipping, so its mean rests on a smaller sample. Figure 5.3 shows the distributions.

Self-reported skin color, socio-economic category, and ethnicity are shown in Figures 5.4 to 5.6, each over the 10,208 participants who answered. For skin color, 67.3% selected the lightest of the three categories, 27.2% the second and 3.3% the third, while 2.3% preferred not to answer. The most frequent socio-economic categories were intellectual and scientific professions at 23.3%, intermediate professions at 18.1%, and executives at 18.0%, followed by retirees at 15.0%. Ethnicity was a multiple-choice question: 76.0% of respondents selected Caucasian, 3.1% Middle-Eastern or North-African, 1.5% Afro-American or Black, 1.2% Asian, 0.7% Amerindian or Alaska Native, and 0.1% Native Hawaiian or Pacific Islander. Among all respondents, 15.4% selected the *other* category and 3.3% preferred not to answer.

Metric	n	p5	Median	Mean	SD	p95
Height (cm)	10,038	158	174	174	9	187
Weight (kg)	8,144	56	79	80	17	110
Age (years)	10,126	29	52	52	14	76
BMI (kg/m^2)	8,121	20	26	27	5	36

Table 5.2: Continuous demographic variables across the analyzable sample, each over the participants who reported a plausible value.

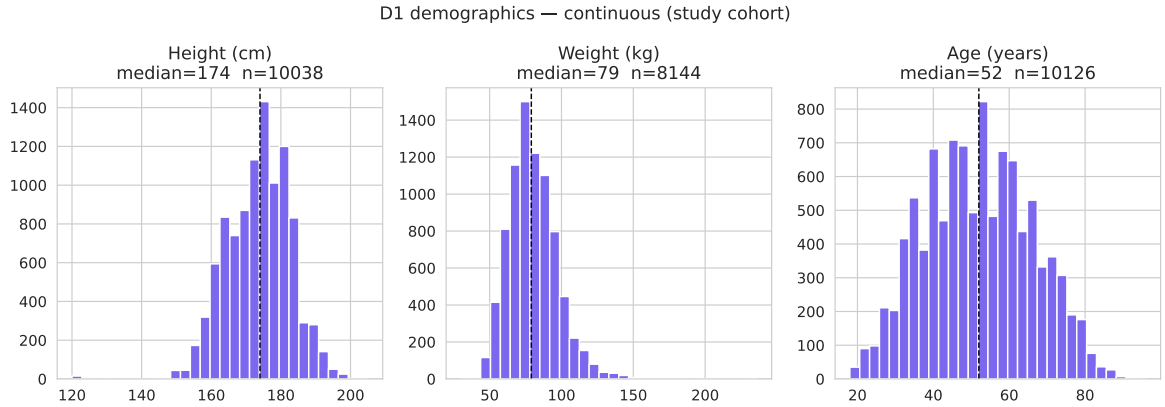


Figure 5.3: Distributions of height, weight, and age across the analyzable sample.

5.3 Definition of the insomnia labels

Two complementary labels describe each participant's insomnia: a diagnosis class derived from DSM-5-style questions, and a severity tier derived from the ISI score.

5.3.1 Insomnia diagnosis classes

The insomnia status of each participant is derived from the *Labeling* part of the survey (section E in Figure 5.1), administered at three iterations across the monitoring period (see Figure 1 of Chapter 4), following the DSM-5 insomnia criteria. A participant who selects only the no-complaint option, reporting neither insufficient quantity nor poor quality of sleep, is assigned good sleep. A participant who reports a complaint is then required to meet the DSM-5 symptom and impairment criteria: at least one difficulty with falling asleep, staying asleep or waking too early (Criterion A), and at least one associated daytime consequence (Criterion B). A complaint meeting neither is labeled symptomatic not insomnia. A complaint meeting both is screened for the differential exclusions, and a participant reporting a co-occurring other sleep disorder, a mental disorder, or a substance cause is labeled secondary insomnia. The remaining participants hold primary insomnia, split by chronicity into chronic insomnia when the difficulties have occurred at least three days per week over at least three months and acute insomnia otherwise. These five classes are good sleep,

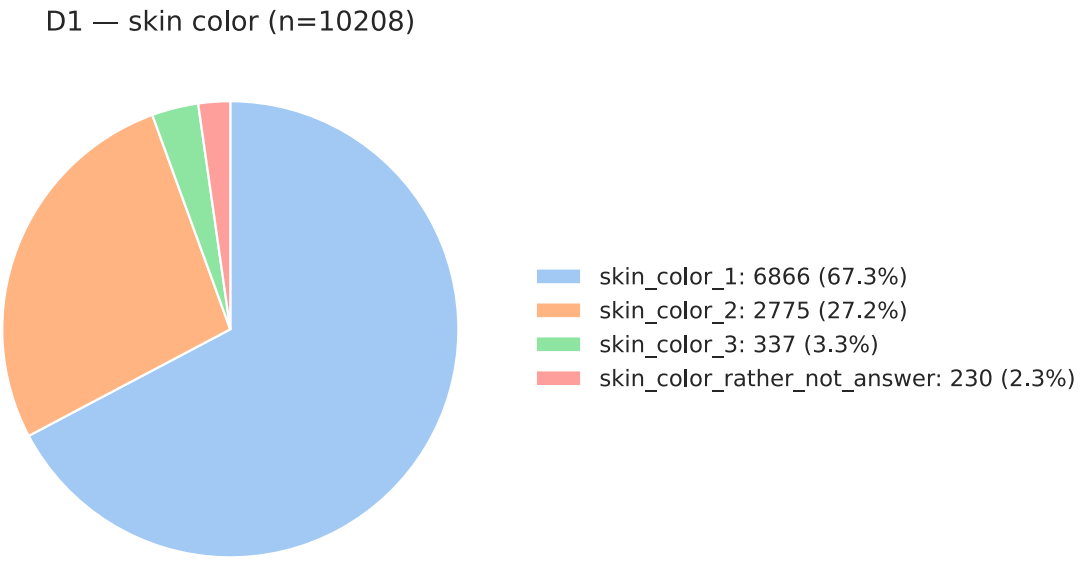


Figure 5.4: Self-reported skin color across the analyzable sample. Category 1 is light, white or fair skin that always burns and never or only sometimes tans; category 2 is medium, olive or moderate-brown skin that sometimes or rarely burns and always tans; category 3 is brown, dark-brown or black skin that rarely burns and tans more than average.

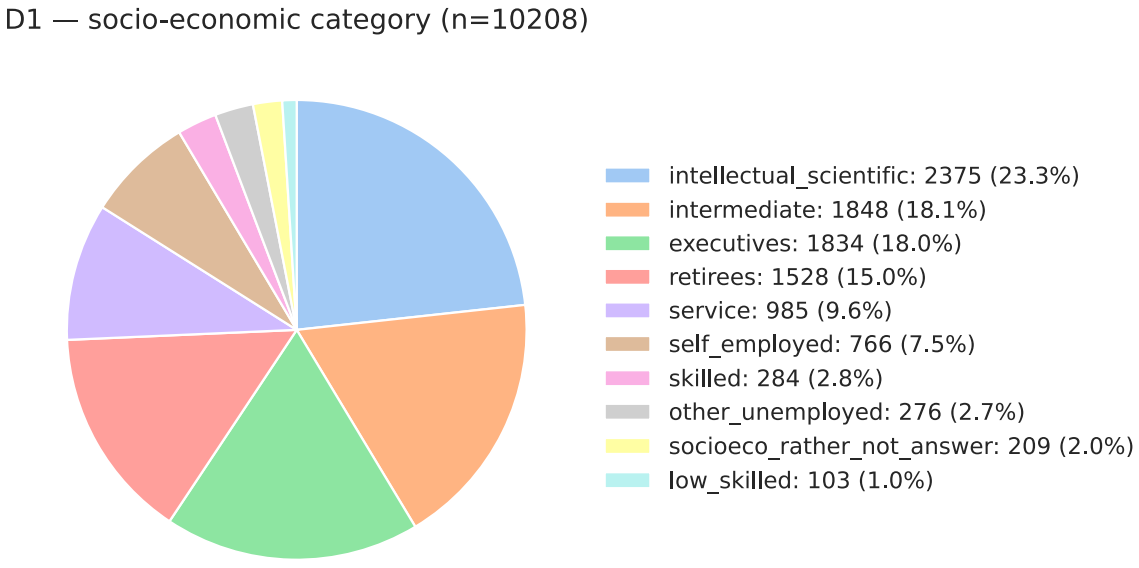


Figure 5.5: Self-reported socio-economic category across the analyzable sample.

symptomatic not insomnia, secondary insomnia, acute insomnia and chronic insomnia.

A finer subtype refines this split by the nature of the complaint. The complaint question records dissatisfaction with sleep quantity, with sleep quality, or with both at once, so

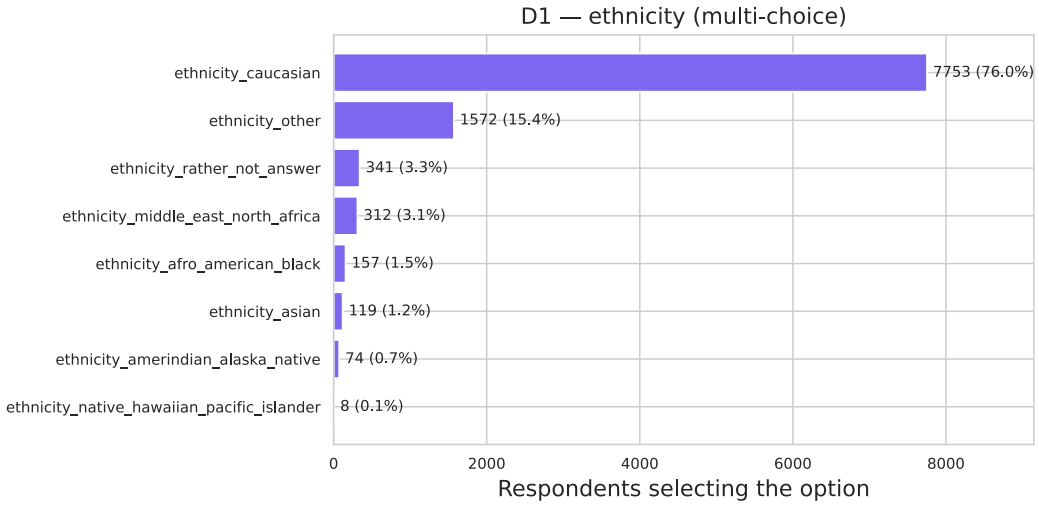


Figure 5.6: Self-reported ethnicity, a multiple-choice question, across the analyzable sample.

each of the primary acute and chronic groups divides into a quantity, a quality and a both subtype. The nine classes are therefore good sleep, symptomatic not insomnia, secondary insomnia, acute quantity, acute quality, acute both, chronic quantity, chronic quality and chronic both.

Non-response rises across the three iterations, from 9.3% of the cohort at the first to 30.4% and 47.8% at the second and third. A fixed-cohort denominator would therefore make every class appear to decline through attrition alone, so prevalence is reported over responders in two ways: cross-sectionally over the respondents at each iteration, and longitudinally over the participants who answered all three. These are shown in Figure 5.7.

Chronic insomnia and good sleep are the most frequent classes, chronic insomnia at 25 to 29% of responders and good sleep at 27 to 33%, followed by secondary insomnia near 25%, symptomatic not insomnia at 10 to 12%, and acute insomnia near 6%; once non-response is accounted for, the prevalences are stable across iterations.

5.3.2 Insomnia severity

Insomnia severity is measured by the ISI, scored out of 28. Because the questionnaire was re-administered every two weeks, each participant's ISI score in the analyses that follow is the mean of their answered iterations, which summarizes severity over the monitoring period rather than at a single moment. This mean is banded into the standard clinical severity tiers: none for 0 to 7, subthreshold for 8 to 14, moderate for 15 to 21, and severe for 22 to 28. These tiers are the severity grouping used throughout the chapter, distinct from the diagnosis classes above, which rest on the complaint and its chronicity.

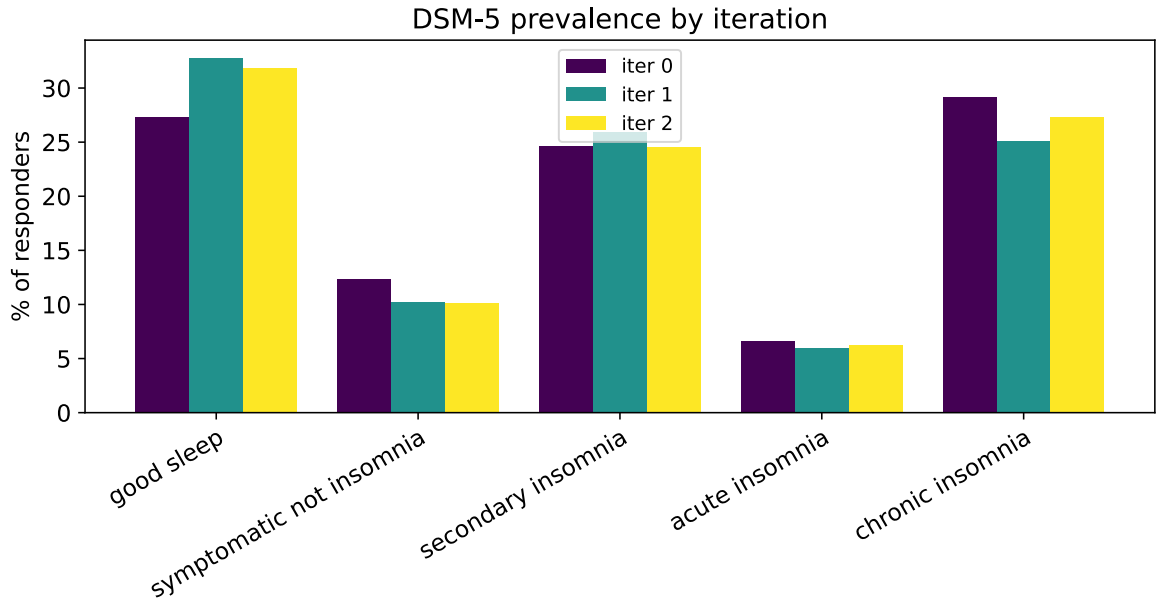


Figure 5.7: DSM-5 insomnia-diagnosis prevalence by survey iteration, over responders: per-iteration responders on the left and consistent responders, who answered all three iterations, on the right.

5.4 Subjective clinical data

5.4.1 Descriptive statistics

The distributions of the subjective clinical scores across the cohort are summarized in Table 5.3. The mean per-user ISI score was 9.2 out of 28, and the PHQ-9 and GAD-7 screens averaged 4.8 out of 27 and 3.5 out of 21 respectively.

Score	n	p5	Mean	SD	p95
ISI (per-user mean)	10,281	1.6	9.2	5.0	18
PHQ-9	10,034	0	4.8	4.1	13
GAD-7	9,974	0	3.5	3.8	11

Table 5.3: Distribution of the subjective clinical scores across the analyzable cohort. p5 and p95 are the 5th and 95th percentiles; SD is the standard deviation; n is the number of cohort users with the score available.

5.4.2 Correlations

The associations in this section are quantified with rank-based methods, since the three self-report scores are bounded ordinal sums with skewed distributions. Pairwise association is

measured by the Spearman rank correlation ρ, which captures any monotonic relationship and is robust to skew and to outlying scores, with a 95% confidence interval from the Fisher z transformation; values near 0.1, 0.3 and 0.5 are conventionally read as weak, moderate and strong. Because insomnia, depression and anxiety are comorbid, partial Spearman correlations are also reported, each holding the third score fixed, so that a coefficient which shrinks towards zero reflects variance shared with that third score whereas one that survives is specific to the pair.

Differences in ISI across the DSM-5 classes are tested with the Kruskal–Wallis test, the rank analogue of one-way analysis of variance, which compares the classes without assuming normality or equal variances; its effect size is epsilon-squared, bounded between 0 and 1, for which 0.01, 0.06 and 0.14 are conventionally small, medium and large. A significant test is followed by Dunn pairwise comparisons with the Holm correction, and the magnitude of each difference is given by the Hodges–Lehmann estimator, the median of all paired between-class score differences, which expresses the typical gap between two classes directly in ISI points, so a larger value is a wider gap. Overlap between classes is summarized by the overlapping coefficient, the shared area of their two score densities, which runs from 0 when the distributions are disjoint to 1 when they coincide, so a higher value means the classes separate less, alongside the overlap of their interquartile ranges.

The association between the ISI clinical severity tiers and the complaint-based class is summarized by Cramér's V, which runs from 0 for no association to 1 for a complete one.

The ISI correlates with the PHQ-9 depression score at $\rho = 0.613$ and with the GAD-7 anxiety score at $\rho = 0.508$, while the depression and anxiety scores correlate with each other at $\rho = 0.707$, each at $p < 0.001$ over the participants who completed both instruments, as Table 5.4 reports. Partial correlations that hold the third score fixed separate the shared association from the specific one: the insomnia–depression coefficient falls to 0.418 whereas the insomnia–anxiety coefficient falls to 0.134, so the link between insomnia and anxiety is largely carried by their common variance with depression, while a substantial insomnia–depression association remains after adjustment.

Pair	n	ρ [95% CI]	Partial ρ [95% CI]
ISI–PHQ-9	9,914	0.613 [0.601, 0.625]	0.418 [0.402, 0.434]
ISI–GAD-7	9,888	0.508 [0.493, 0.522]	0.134 [0.114, 0.153]
PHQ-9–GAD-7	9,974	0.707 [0.697, 0.717]	0.579 [0.566, 0.592]

Table 5.4: Spearman correlations among the subjective scores, each computed over the participants who completed both questionnaires of the pair, so n varies. The partial coefficient controls for the remaining score over the $9{,}888$ participants with all three; all $p < 0.001$.

Insomnia severity scales with the DSM-5 class. Median ISI is 4.5 in good sleepers, 6.8 in the symptomatic-not-insomnia class, 8.0 in acute insomnia and 10.9 in chronic insomnia, and reaches its highest value of 12.3 in secondary insomnia, above the primary chronic class; the Kruskal–Wallis effect size is large, with $\varepsilon^2 = 0.333$ at $p < 0.001$ and every pairwise comparison significant after Holm correction; the Hodges–Lehmann median shift reaches 7.1 ISI points between secondary insomnia and good sleep and 5.8 points between

primary chronic insomnia and good sleep, as Table 5.5 reports. Cross-tabulating the ISI clinical severity tiers against the complaint class gives a Cramér's V of 0.315, so questionnaire severity and the complaint-based class are related but discordant. The adjacent classes nonetheless overlap on the ISI axis, sharing a mean overlapping coefficient of 0.80 between neighboring pairs. The finer subtype split shows the same graded pattern with a marginally larger effect size of $\varepsilon^2 = 0.341$, as Figure 5.8 shows.

Class	n	Median ISI	IQR
Good sleep	2,684	4.5	2.4–7.7
Symptomatic not insomnia	1,209	6.8	4.5–9.6
Secondary insomnia	2,425	12.3	9.0–15.1
Acute insomnia	650	8.0	5.6–10.5
Chronic insomnia	2,865	10.9	8.0–13.9

Table 5.5: ISI score by DSM-5 insomnia class at the first labeling iteration. IQR is the interquartile range. Denominators are the class members with an ISI score available and sum to $9{,}833$; the comorbidity table (Table 5.6) instead reports full class size and sums to the $9{,}866$ first-iteration responders.

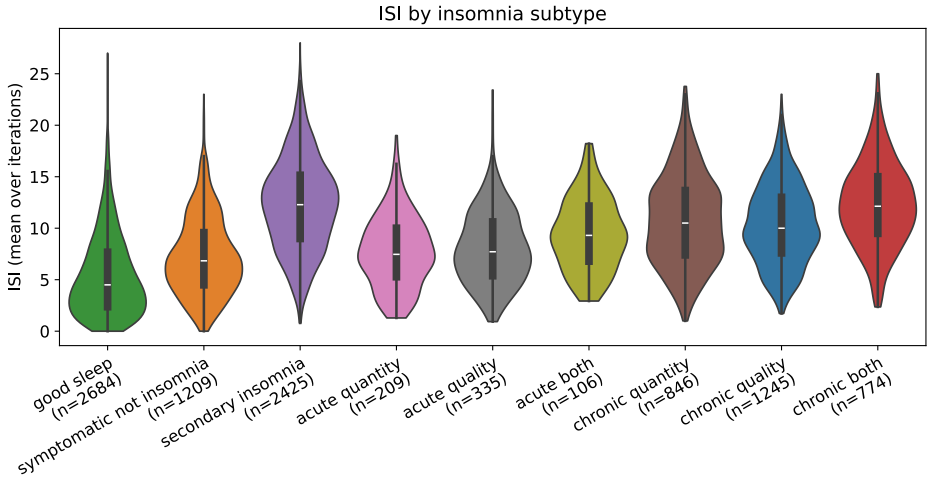


Figure 5.8: ISI distribution by the insomnia subtype at the first labeling iteration. The inner marker is the median and the box the interquartile range.

Using the baseline ISI rather than its mean, the score and the complaint-based class agree only partly, with a tier-by-class Cramér's V of 0.346. The disagreement is asymmetric: 2,031 participants report a DSM-5 complaint while their baseline ISI sits in the lowest tier, whereas only 7 report good sleep at a severe baseline ISI, with 474 reporting good sleep at a raised ISI. This low-severity, high-complaint group is not uniform: 753 of the 2,031 belong to the symptomatic-not-insomnia class, whose complaint does not by itself meet the insomnia criteria, so a low score is expected, while the rest meet an insomnia class yet rate their sleep only mildly disturbed, since the DSM-5 label records whether a qualifying complaint with daytime impairment is present whereas the ISI grades how severe it feels.

Across the nine-class subtype, only secondary insomnia and the chronic both-complaint subtype share an indistinguishable baseline ISI distribution; the other pairs differ in median. Good sleep sits at the lowest median, yet the class distributions overlap enough that, despite this graded ordering, the score does not cleanly separate the complaint classes at the individual level. The ISI is a validated self-report outcome measure that quantifies the perceived severity of insomnia and its daytime impact [22]; as a perception-based instrument it indexes the subjective complaint rather than objective sleep, so its limited agreement with the criteria-based label and with device-measured sleep reflects a difference in what each measure captures rather than a defect in either.

5.4.3 Comorbidity

Depression and anxiety, screened by the standard moderate-or-worse cut-offs of PHQ-9 ≥ 10 and GAD-7 ≥ 10, become more frequent as insomnia worsens. Across the DSM-5 classes the prevalence of depression is 4.8% in good sleepers, 7.3% in the symptomatic-not-insomnia class, 3.5% in acute insomnia and 13.9% in chronic insomnia, and is highest at 22.7% in secondary insomnia; anxiety follows the same ordering, from 2.9% in good sleepers to 14.2% in secondary insomnia, as Table 5.6 reports. Graded by ISI severity tier the same prevalences span 2.0% to 68.2% for depression and 1.8% to 53.6% for anxiety from the lowest to the highest tier, shown in Appendix Figure D.1. Among the $9{,}888$ participants with all three scores, moderate insomnia defined as ISI ≥ 15 co-occurs with both depression and anxiety in 2.7%, whereas 7.5% report this level of insomnia without either depression or anxiety, so insomnia is frequently but not exclusively accompanied by depression or anxiety symptoms, as Appendix Figure D.2 details.

Class	n	PHQ-9 ≥ 10	GAD-7 ≥ 10
Good sleep	2,693	4.8%	2.9%
Symptomatic not insomnia	1,213	7.3%	4.5%
Secondary insomnia	2,433	22.7%	14.2%
Acute insomnia	651	3.5%	2.3%
Chronic insomnia	2,876	13.9%	8.2%

Table 5.6: Depression (PHQ-9 ≥ 10) and anxiety (GAD-7 ≥ 10) prevalence by DSM-5 insomnia class.

5.5 Objective data

5.5.1 Data segmentation

The objective features are computed over temporal windows aligned to the recurring questionnaire. Each participant's monitoring period is split at the timestamps of the fourteen

ISI submissions, which act as internal boundaries and therefore define up to fifteen windows. The first window runs from consent to the first submission and may be too short to use, in which case it is treated as an artefact and left aside. A window wider than fifteen days is cropped to its final fifteen, because the ISI answer is taken to represent roughly the participant's preceding two weeks of sleep, the recall horizon illustrated in Figures 1 and 2 of Chapter 4, so that earlier and forgotten nights do not enter the window. Within each window the per-night sleep and actigraphy features are computed, summarized per window, and reduced to one value per participant by the subject-level median, the unit that the feature description and the phenotyping both use.

A participant enters the objective cohort when consent is recorded without a later exclusion or withdrawal, at least one window qualifies, and the cleaned matrix retains the participant. A window qualifies only if it comes from a SW, carries firmware version 2941, and contains at least nine worn nights, so that the actigraphy and sleep signals are complete and comparable across participants. The subject matrix is then cleaned by dropping features and subjects with more than 40% missing values and imputing the few remaining gaps. The imputation fills each missing value from the five nearest subjects weighted by inverse distance, so that closer subjects contribute more, with the distance taken as the Euclidean distance over the features that two subjects share; to keep this distance from being dominated by large-magnitude features, it is computed on z-standardized features and the imputed values are then returned to their original units. Of the 11,872 users processed, 9,190 were eligible with a qualifying window, and 8,738 remained after the cleaning.

This cohort of 8,738 subjects sits between the two populations used elsewhere in the chapter. It is larger than the 5,478 participants who met the data-contribution criteria reported earlier, since the present gate requires only a qualifying device window rather than usage over more than 60% of the study, and it is smaller than the 10,878-participant analyzable study cohort, since not every consenting participant contributed a qualifying window. The two-window requirement applies only to the reliability estimate reported below, not to cohort inclusion, so one qualifying window admits a participant. Device coverage does not fall as insomnia severity increases, so the objective comparisons are not biased by participants with more severe insomnia wearing the devices less, as Appendix Figure D.3 shows.

5.5.2 Feature building

The objective feature set is the one fixed for the phenotyping analysis: thirty-one subject-level median summaries of the sleep, cardiac and activity measures, together with 8 nonparametric circadian metrics computed once per subject over the whole window, 39 features in all, each defined in Appendix Table D.2. No feature exceeds the 40% missingness threshold, so the cleaned matrix that the feature description and the clustering use keeps all 39 features over the 8,738 retained subjects. Each median feature is summarized per day, then per window, then reduced to the subject median, and its paired within-person standard deviation across nights measures night-to-night variability.

Table 5.7 summarizes the main features. Median TST is 7.32 hours and median SE 0.97, the total awake duration has a median of 12.4 minutes, and the rest-activity rhythm is mod-

erately stable with an interdaily stability of 0.33. Before imputation, feature-wise missingness is at or below 17.3%. Night-to-night variability is substantial, with a median within-person standard deviation of 1.12 hours for TST. The distributions of the main features are shown in Figure 5.9, and their correlation structure, missingness and circadian detail in Appendix Figures D.4, D.5 and D.6.

Feature	n	Median	p5–p95	Within-person SD
Total sleep time (h)	8,737	7.32	5.65–8.65	1.12
Deep-sleep duration (h)	8,737	3.17	1.82–4.69	0.72
Light-sleep duration (h)	8,737	4.02	2.59–5.44	0.89
Time in bed (h)	8,738	7.57	5.91–8.97	1.17
Total awake duration (min)	8,737	12.42	2.93–30.50	10.08
Sleep efficiency	8,698	0.97	0.93–0.99	0.02
Sleep-maintenance efficiency	8,698	0.98	0.94–1.00	0.02
Awakening index (events/h)	8,698	0.16	0.00–0.46	0.14
Resting heart rate (bpm)	8,723	69.17	54.60–84.00	6.44
Steps (per day)	8,735	5,366	1,850–10,884	2,685
Interdaily stability	7,258	0.33	0.19–0.47	—
Intradaily variability	7,258	0.62	0.44–0.88	—
Relative amplitude	7,258	0.89	0.74–0.95	—

Table 5.7: Distribution of the main objective features across the 8,738-subject cohort. The n column gives the number of subjects for which each feature is available. The last column is the median across subjects of each subject's night-to-night standard deviation; circadian metrics are single per-subject summaries on each window.

Because the feature set is high-dimensional, it is projected by PCA onto its first two components to see whether a two-dimensional view retains most of its variance. It does not: the first two components explain 21.4% and 17.9% of the variance, 39.3% together, so the features are not organized around a few dominant axes, as Figure 5.10 reports, which anticipates the clustering results of the following section.

5.6 Clustering

5.6.1 Model Selection

The cleaned matrix of 39 features over 8,738 subjects is clustered in two ways, a panel of standard partitioning algorithms and a non-linear AE representation, so that a negative result cannot be attributed to a single family of methods.

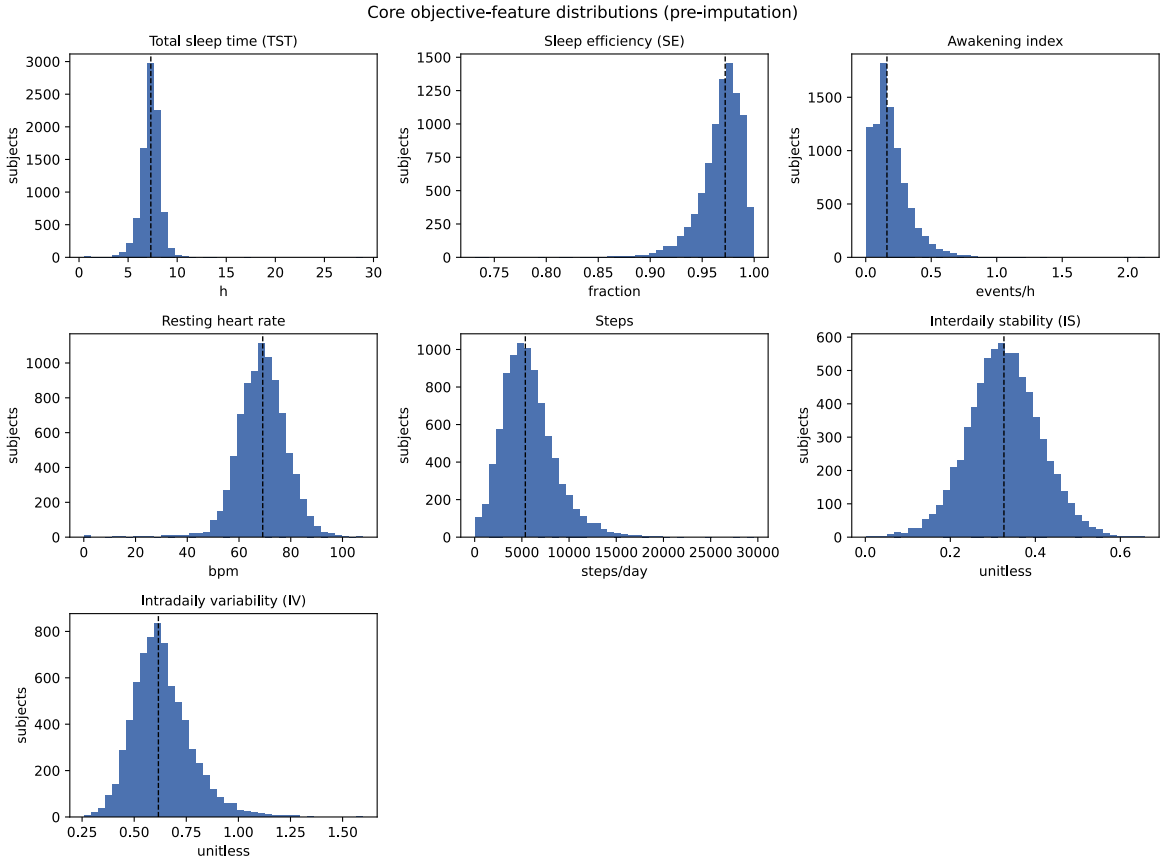


Figure 5.9: Distributions of the main sleep, cardiac, activity and circadian features across the objective-analysis cohort.

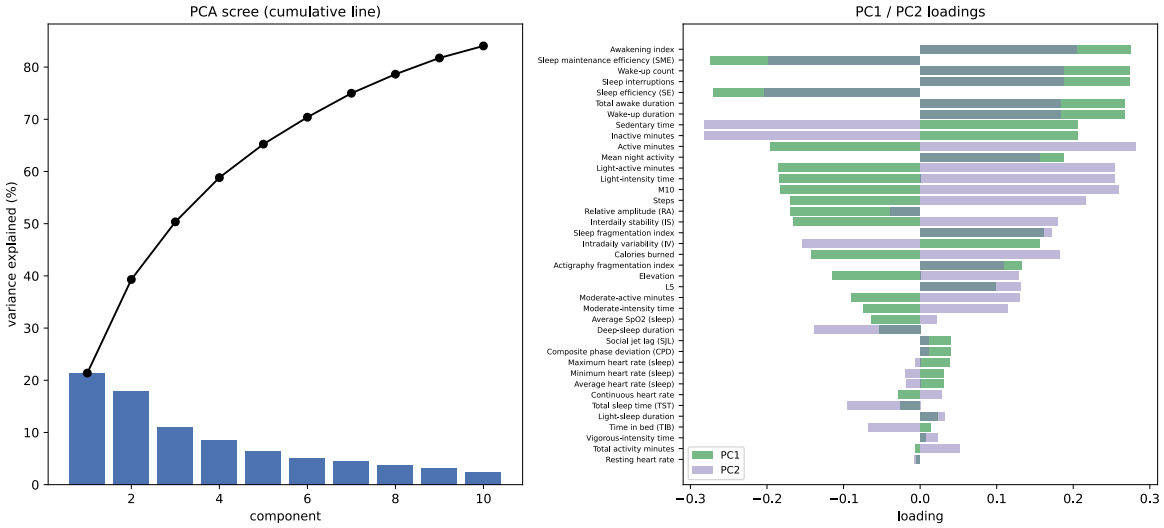


Figure 5.10: PCA of the objective feature set: scree plot and the first two component loadings.

5.6.1.1 Traditional partitioning

The first approach applies a panel of standard algorithms, comprising k-means, Gaussian mixtures and agglomerative hierarchical clustering, each scanned over its hyperparameters with the number of clusters chosen to maximize the silhouette, together with the density-based HDBSCAN and mean shift algorithms. Every candidate partition is assessed for internal validity by the silhouette together with the Davies–Bouldin, Calinski–Harabasz and Dunn indices, and for agreement with the clinical labels, the DSM-5 insomnia class and the ISI severity tier, by the adjusted Rand index. Linear discriminant analysis measures how separable the clinical labels are within the feature space, and projections by PCA and Uniform Manifold Approximation and Projection (UMAP) give the visual counterpart. Because the silhouette rewards any compact partition, including an arbitrary slice of a continuous cloud, a partition is judged by its agreement with the clinical labels, by the density-based validation and by the label separability.

5.6.1.2 Non-linear autoencoder

The second approach tests whether a non-linear representation exposes structure that the panel, working in the original feature space, might miss. An AE maps the 39 standardized features through hidden layers of width 128 and 64 to an 8-dimensional latent and is trained for 60 epochs by reconstruction, and the latent means are then clustered with HDBSCAN at a minimum cluster size of $\max(15,\ 1\%\ \text{of}\ n)$, where n is the number of subjects clustered.

The clustering pipeline was itself checked on a synthetic control built at the same 39-feature width. On data with a single continuous regime it found no clusters across all five random seeds, and on data with three density-separable regimes it recovered exactly three across all five seeds, at a mean adjusted Rand index of 0.985 against the injected labels. The pipeline therefore finds density-separable structure when it is present and reports none when it is absent. Its limit is that a subgroup that is real but not set apart by a density gap in this feature space would be missed, by the partitioning panel and the AE alike.

5.6.2 Results

5.6.2.1 Traditional partitioning

Clustering the objective features does not recover discrete phenotypes. The best partition that does not merely isolate a few outliers, a two-group hierarchical solution in which each cluster holds at least 5% of the cohort, reaches a silhouette of only 0.21, and its adjusted Rand index is 0.014 against the ISI severity tier and -0.001 against the DSM-5 class, both indistinguishable from chance; the partitions that score a higher silhouette do so only by splitting off a negligible outlier group. The low-dimensional projections show a single diffuse cloud rather than separable groups, as Figure 5.11 shows, with the internal-validity scores across the algorithms given in Appendix Figure D.7. The profile of this best partition confirms its weakness directly: the two groups depart only slightly from the cohort mean

on every feature and strongly on none, so the split is a shallow cut of a continuous distribution rather than a contrast between distinct sleep-activity types, as Appendix Figure D.8 shows.



Figure 5.11: PCA (top) and UMAP (bottom) projections of the objective feature matrix, with each DSM-5 class highlighted in color over the full cohort in gray. The classes occupy a single diffuse cloud rather than separable regions. The highlighted classes are the five DSM-5 diagnosis classes.

Coloring the same projection by individual feature values rather than by clinical label confirms why no partition succeeds. The objective features vary smoothly across the embedding, with clear gradients along the fragmentation and daytime-activity axes, yet the mean ISI score tracks these axes only weakly, its Pearson correlation reaching 0.14 with the first principal component, the fragmentation axis, and 0.10 with the second, the daytime-activity axis. The structure the features carry therefore runs largely orthogonal to insomnia severity, as Figure 5.12 shows, so a grouping built on the dominant feature variation would separate participants by something other than their insomnia.

5.6.2.2 Non-linear autoencoder

The AE reaches the same conclusion. At the end of the 60-epoch training its reconstruction captured 87.8% of the variance of the standardized features, the share of input variance the decoded output reproduces, which rose only to 89.8% when training was extended to 200 epochs, so the latent is not undertrained. All 8 latent dimensions were in use at per-dimension standard deviations between 1.47 and 2.11, so the latent did not collapse. Clustering this latent with HDBSCAN returned no clusters at every minimum cluster size of ten or more, on the full 8,738 subjects, labelling the entire sample as noise. The only setting that produced any groups was a minimum cluster size of five, which split the cohort into one large group of 7,804 subjects and three negligible groups of 9, 5 and 5 subjects.

That single partition results in a low adjusted Rand index of 2×10^{-4} against both the severity tier and the diagnosis, both at chance, and the density-based clustering validation index, which rewards partitions whose clusters are internally dense and well separated and runs from -1 to 1, is -0.295 on the standardized original matrix, the negative value

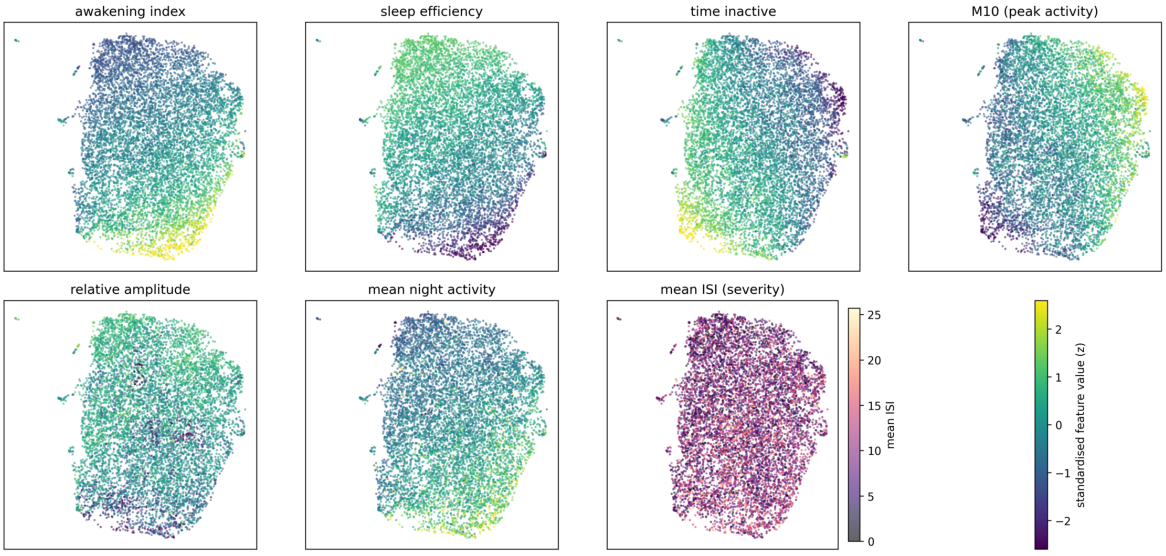


Figure 5.12: UMAP projection of the objective feature matrix colored by six standardized feature values and, in the final panel, by the mean ISI score. The features form smooth gradients along the fragmentation and activity axes, while insomnia severity tracks them only weakly (Pearson $|r| \leq 0.15$).

indicating no real density gap. The partition is therefore orthogonal to the clinical labels and density-negative, a fragmentation of a continuous distribution rather than phenotype structure; the projected latent shows the severity tiers fully intermixed in a single diffuse cloud, as Figure 5.13 shows.

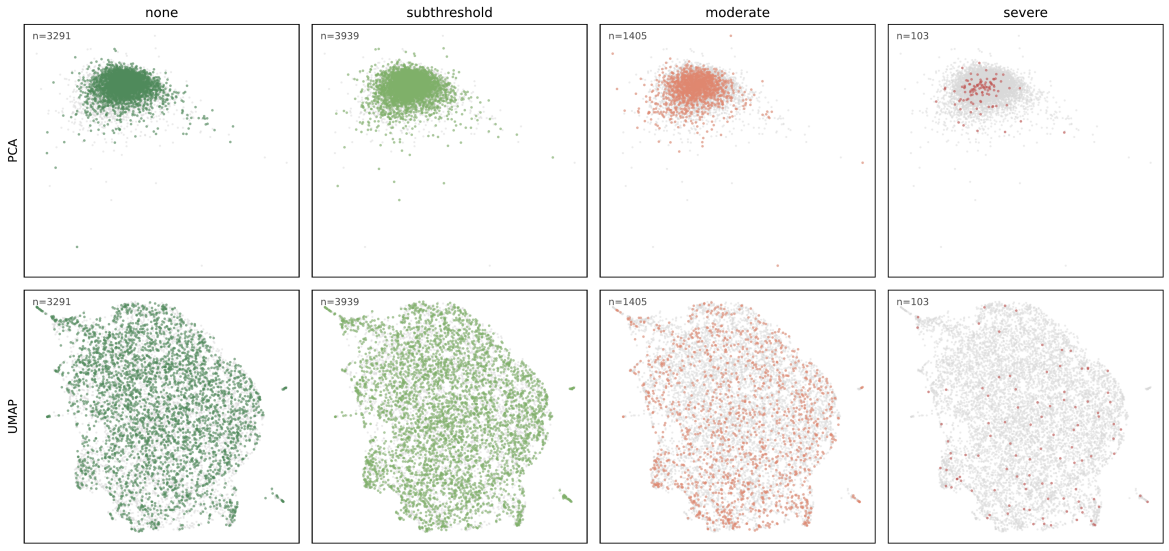


Figure 5.13: PCA (top) and UMAP (bottom) projections of the autoencoder latent, with each ISI severity tier highlighted in color over the full cohort in gray. The tiers occupy a single diffuse cloud rather than separable regions.

5.6.2.3 Interpretation

Both approaches run on the identical matrix of 8,738 subjects and 39 features and agree. The partitioning methods returned partitions only because most of them are required to cut the data into a preset number of groups, and the silhouette they maximize rewards any compact slice of a continuous cloud, whereas the density-based validation and the near-zero label agreement reported the absence of real groups. The density-based clustering skips the forced-partition step and reports the noise directly, and the AE shows that a non-linear representation does not create a density gap that the original space lacked. Discrete, separable insomnia phenotypes are therefore absent in these aggregate features under both linear and non-linear methods and under a detector shown able to recover structure when it exists, which answers the fourth research question in the negative for discrete subtypes and reframes the objective sleep-activity phenotype as a continuum.

This result limits what can be claimed from it. The lack of a density gap rules out discrete insomnia subtypes, but it does not forbid splitting the cohort into convenient strata such as severity bands or risk groups. Such strata would not be phenotypes: their boundaries are arbitrary and move with the number of groups chosen. They would also not track insomnia, because the directions along which these features vary most are largely unrelated to the insomnia labels, so grouping by them would sort participants by something other than insomnia, while grouping by severity would only restate the ISI score. A separate question, not settled here, is whether the per-subject aggregation of the features, rather than the choice of encoder, is what leaves no separable signal; the temporal analyses discussed later are aimed at it.

5.6.2.4 Night-to-night variability

The phenotyping so far described each participant by the median level of every feature. A supplementary analysis asked whether night-to-night variability, which the median discards, carries structure the level misses. Two summaries of the same recordings were clustered: the raw night-to-night variability of each feature, measured as the root mean square of its successive differences, and a residual variability that keeps only the part of that fluctuation not explained by the participant's own level or by how many nights they contributed, so that it reflects instability rather than how much or how well they sleep. Each was clustered as before and then judged in two ways. The first compares it against clusterings of shuffled copies of the data, in which every feature's values are reassigned at random across participants, so that each feature keeps its own distribution while the way values combine within a participant, and with it any real grouping, is destroyed; a genuine partition should be tighter than anything the shuffled copies yield. The second redraws the sample repeatedly and asks whether the same split recurs.

Neither variability summary yields real clusters: their silhouettes stay below 0.18, no higher than the shuffled copies produce, and the raw summary does not recur when the sample is redrawn. Night-to-night variability therefore carries no discrete phenotype beyond the level continuum. Two cautions temper this. The comparison against shuffled copies is demanding, because shuffling each feature on its own occasionally lines up a few

extreme values into a small tight cluster that raises the silhouette the real partition must beat, so the conclusion rests more on how small the silhouettes are than on that comparison alone. And the residual summary's split does recur under resampling, yet recurrence is not evidence of a real boundary, since a single continuous cloud can be cut in the same place every time.

5.6.2.5 Beyond a cross-sectional partition

A cross-sectional clustering uses one value per participant, so it cannot see how a participant changes over time; two within-person measures fill that gap. The first is night-to-night variability. The objective features fluctuate within a participant from one night to the next, for instance total sleep time varies by a within-person standard deviation of 1.12 hours, as Table 5.7 reports; and for the ISI, an intraclass correlation of 0.75 means that about one quarter of the variation in scores comes from a participant changing across the twenty-six weeks rather than from differences between participants, as Figure 5.14 shows. The second is how severity moves over follow-up. Across all pairs of consecutive ISI answers, a participant stays in the same severity tier with an average probability of 0.64, ranging from 0.81 in the none tier down to 0.48 in the severe tier, and, when the tier does change, almost always moves to an adjacent one, so most participants hold their tier from one two-week assessment to the next while a minority improve or worsen, as Figure 5.15 shows. Both measures describe within-participant change, which a single-snapshot clustering cannot represent.

5.7 Subjective–objective relationship

Having described the subjective and the objective measures separately, this section relates them directly over the 8,738 subjects who carry both. The screen is comprehensive: every available subject-level device feature, not only the 39 used for clustering, is correlated with each of the three self-report scores, the ISI, the PHQ-9 and the GAD-7, by the Spearman coefficient, with Benjamini–Hochberg false-discovery control across the many tests. The ISI associations are additionally computed as partial correlations that hold the PHQ-9 and GAD-7 scores fixed, isolating the insomnia-specific signal. The screen is descriptive and implies no causation.

The associations are weak in magnitude despite reaching significance, which the large sample makes unsurprising. Of the features tested, 59 reach significance at a q-value below 0.05, and the strongest absolute coefficients are near 0.19, led by Relative Amplitude (RA) and the L5 circadian metric against the ISI and by composite phase deviation against the PHQ-9, as the full screen in Appendix Figure D.9 reports. Of the 59 significant features, 38 are among the 39 fixed objective features used for clustering and 21 are not. Group comparisons across the ISI severity tiers and the DSM-5 classes are statistically significant but explain very little of the variance: the largest effect size is an ε^2 of 0.028, about 3% of the rank variance and well below the 0.06 medium-effect mark, as Appendix Figures D.10 and D.11 detail.

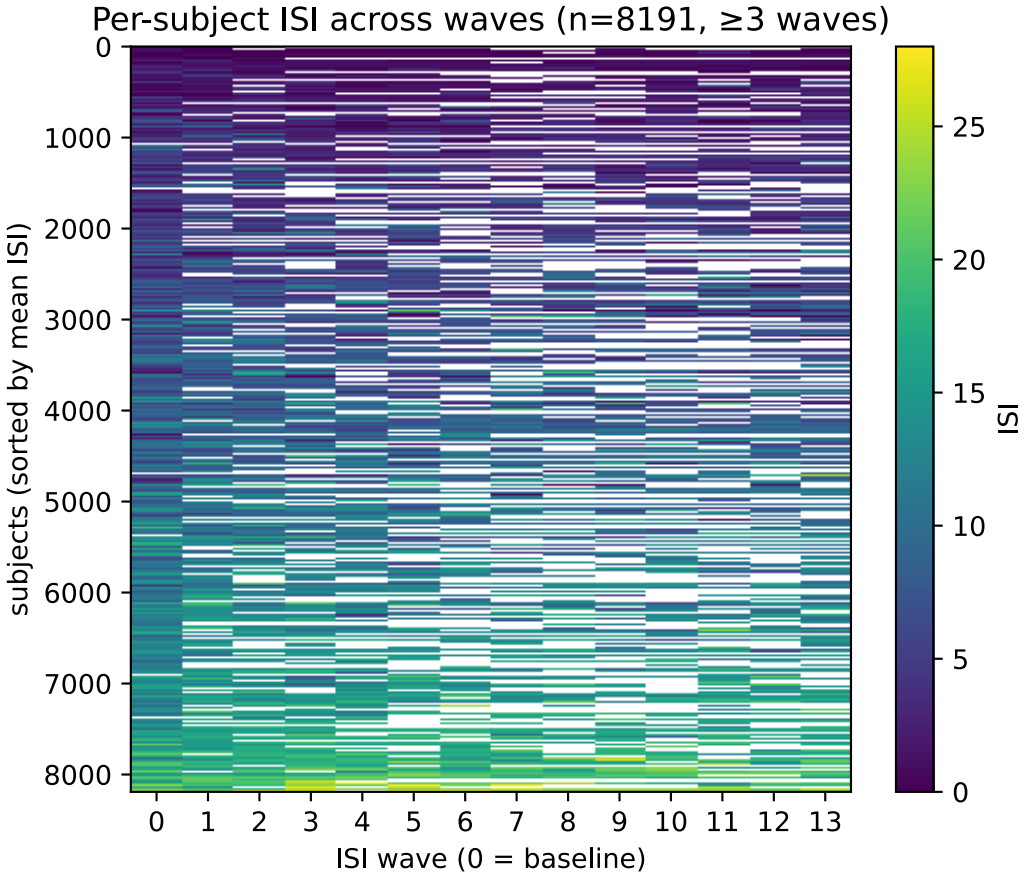


Figure 5.14: Per-participant ISI trajectories across the biweekly administrations, displayed as a heatmap with one row per participant ordered by mean score, showing within-person fluctuation around a largely stable level.

Holding the depression and anxiety scores fixed leaves a small insomnia-specific subset whose partial coefficients remain near 0.16, led by the awakening index at 0.171, mean night activity at 0.166, the L5 metric at 0.163, sleep efficiency at -0.160, RA at -0.159 and sleep-maintenance efficiency at -0.159, as Figure 5.16 shows. This residual signal concentrates in fragmentation and circadian features rather than in sleep duration. Night-to-night variability tracks the scores at a similar weak magnitude, for instance the awakening-index standard deviation against the ISI at 0.173, detailed in Appendix Figure D.12. Because no self-reported sleep-duration estimate is collected, the classic sleep-state misperception comparison is unavailable, so a proxy across the self-reported complaint dimension is given in Appendix Figure D.13.

A minimum-redundancy ranking provides a parsimony check on the screen: it orders the features by their information about each score while penalizing redundancy with the features already chosen, so a small non-redundant subset carrying the score-related information would surface if one existed, as Appendix Figure D.14 reports. It broadly agrees with the correlation screen at the top of the ranking, which shows the weak associations are not an artefact of redundant features rather than revealing any stronger combined signal; because it uses the score to choose features, it is a descriptive ordering, not a validated

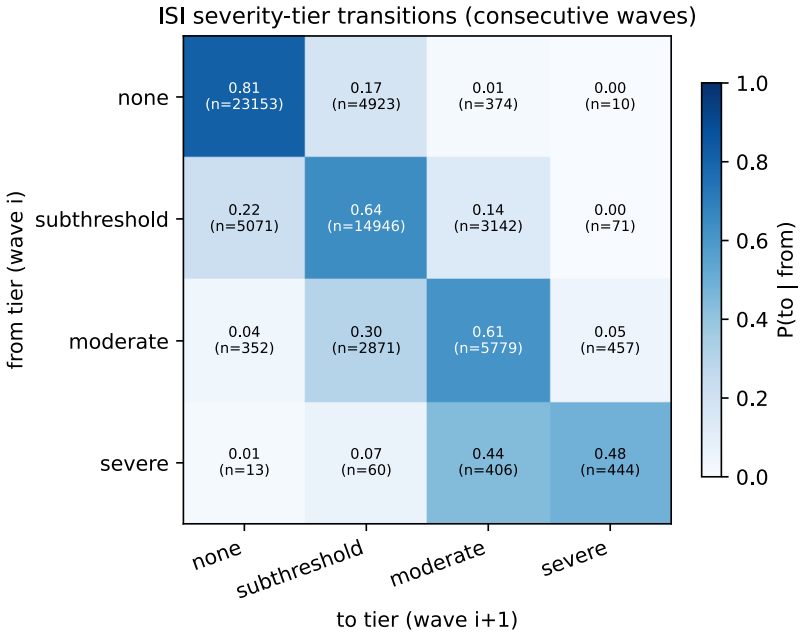


Figure 5.15: Wave-to-wave transitions between ISI severity tiers, pooled over all consecutive answered pairs. Each row gives the probability of the next tier given the current one.

selector or independent evidence. It re-orders the same weak univariate associations and does not test whether the features predict a score jointly or through interactions: that multivariate hypothesis, that markers each weak on their own might together track insomnia, is left to the supervised analyses outlined in the discussion. Taken together, self-reported insomnia severity is only weakly mirrored in objective sleep architecture, with a small genuinely insomnia-specific component concentrated in fragmentation and circadian markers.

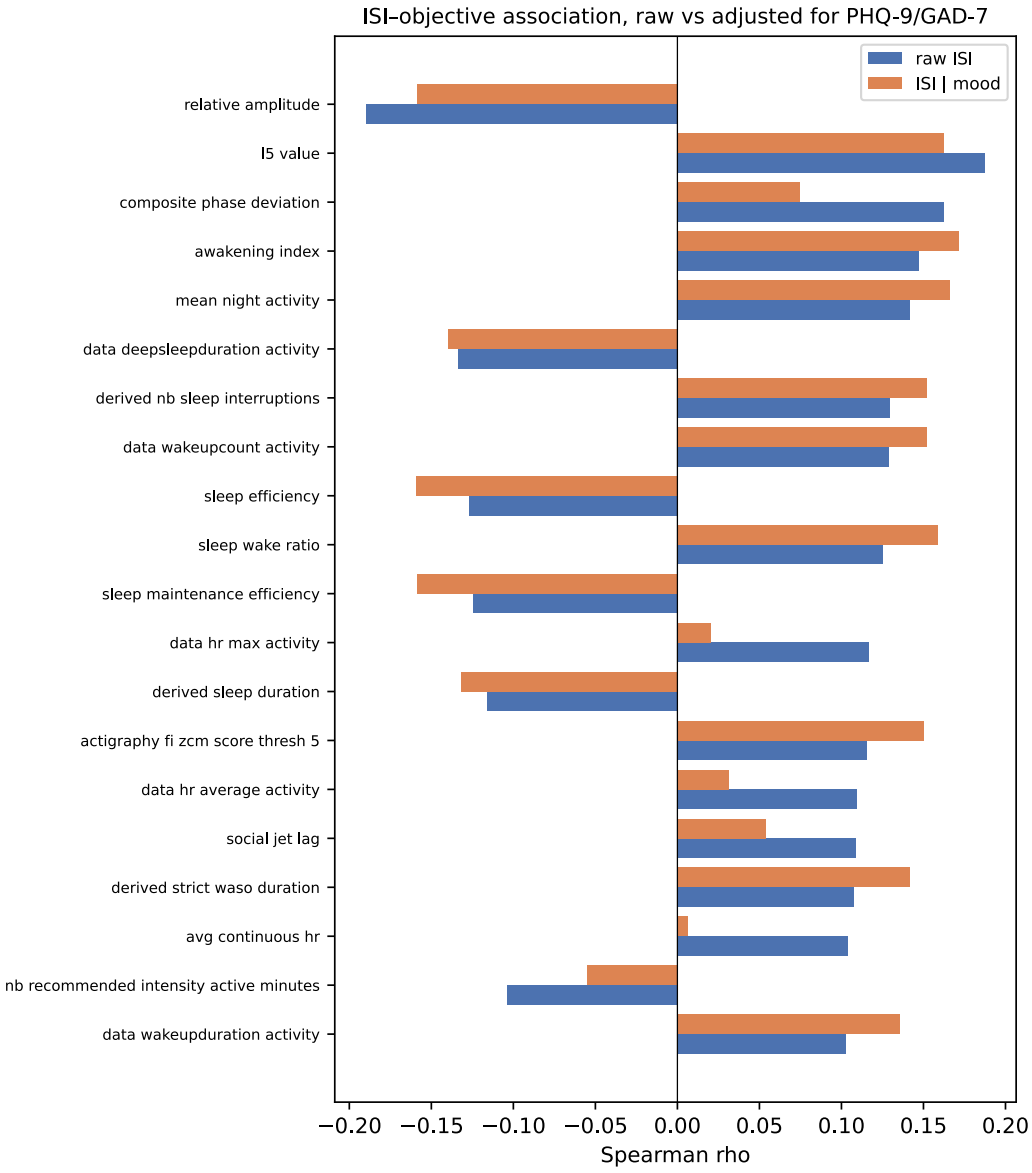


Figure 5.16: Partial Spearman correlation of each feature with the ISI, holding the PHQ-9 and GAD-7 scores fixed, which isolates the insomnia-specific signal.

5.8 Chapter Summary

Table 5.8 gathers the chapter's main results; the points that follow expand each.

Result	Key figures	Reading
Subjective measures		
Score coherence	ISI–PHQ-9 $\rho = 0.61$, ISI–GAD-7 $\rho = 0.51$	Self-report scales agree and cohere
Severity gradient	Median ISI 4.5 in good sleep to 12.3 in secondary insomnia, $\varepsilon^2 = 0.333$	Graded across classes, with substantial overlap
Comorbidity gradient	Depression 4.8 to 22.7%, anxiety 2.9 to 14.2% across classes	Comorbidity scales with insomnia class
Objective sleep–activity features		
Feature space	Median total sleep time 7.32 h, efficiency 0.97, interdaily stability 0.33; first two components 39.3% of variance	High-dimensional, no dominant axis
No discrete phenotypes	Best partition silhouette 0.21, label agreement near zero; AE latent all-noise	A continuum, not discrete subtypes
Link and dynamics		
Subjective–objective link	Strongest device–score $\lvert\rho\rvert \approx 0.19$, group $\varepsilon^2 \leq 0.028$	Objective architecture weakly mirrors severity
Night-to-night variability	ISI ICC 0.75, total-sleep-time within-person SD 1.12 h	Trait-dominant, about one quarter within-person
Severity persistence	Tier-to-tier retention probability 0.64, moves mostly to adjacent tiers	Severity is largely stable over time

Table 5.8: Key results of the SLEEPI analyses. Figures are as reported in the sections above; subjective analyses use the 10,878-participant analyzable cohort and objective analyses the cleaned cohort of 8,738 subjects with usable device data. ρ is the Spearman correlation, ε^2 the Kruskal–Wallis effect size, and the intraclass correlation (ICC) partitions ISI variance into between- and within-person parts.

- Scope: paired subjective and objective measures in the SLEEPI cohort, testing whether the objective sleep-activity phenotype separates into discrete subtypes.
- Comorbidity: all three conditions co-occur in only 2.7% of complete cases.
- No discrete phenotypes: the objective phenotype is better represented as a continuum, ruling out discrete subtypes without excluding a pragmatic severity stratification.
- Subjective–objective link: the weak association carries a small genuinely insomnia-specific component, surviving adjustment for depression and anxiety and concen-

trated in fragmentation and circadian markers.

- ISI severity is persistent: most participants keep their severity tier from one biweekly assessment to the next while a minority remit or worsen.
- Limitations: the clustering is cross-sectional on per-subject aggregates, so it cannot capture within-person change and may miss temporal structure.
- Take-home: insomnia in this cohort is graded rather than categorical; subjective severity is only weakly mirrored in objective sleep architecture; and the within-person night-to-night variability and severity dynamics are the most original direction the dataset opens.

Chapter 6

Discussion

6.1 Reading the findings together

We set out to specify the insomnia phenotype and to identify homogeneous subgroups, from the premise that insomnia is defined by a subjective complaint yet would gain from objective, longitudinal measurement [21]. Each of the four research questions was addressed in its own chapter; here the findings are read together and set against the wider literature. Figure 6.1 gives an overview of the four contributions.

The need for the work follows from the state of the field. The systematic review assembled 49 studies whose outputs fall into four types, and methodological heterogeneity together with weak validation left no basis for a pooled estimate of performance. Multi-night designs were few, shared benchmarks and standardized deployment reporting were absent, and attempts to combine subjective and objective markers were rare. Wearable sleep technology has matured but still wants a standardized way to assess it before its outputs can be trusted across devices and studies [121]. An integrated, longitudinal dataset that records subjective and objective markers in the same individuals therefore had to be built. The SLEEPI protocol did so at scale and with few constraints, monitoring a large cohort at home over 26 weeks under minimal exclusion criteria, which no existing source provided.

One of the sleep-tracking devices used in SLEEPI was evaluated directly against PSG under real-world, free-living home conditions, with participants keeping their usual routines and no restrictions on caffeine or alcohol consumption, screen time, or sleep schedule. The Withings Sleep Analyzer discriminates sleep from wake more reliably than it resolves individual sleep stages, and its agreement is bounded in part by the reference itself, since expert scorers concur only moderately and least of all on light sleep [122, 123, 124]. Its accuracy held across demographic and environmental subgroups and, notably, across self-reported sleep quality [122], a robustness that matters because consumer trackers commonly lose accuracy in disturbed sleep [1]. That no standardized procedure yet exists for evaluating such devices is part of why accuracy is hard to compare across the field [125]. The evaluation nonetheless shows that unobtrusive home sleep staging can be measured with usable accuracy.

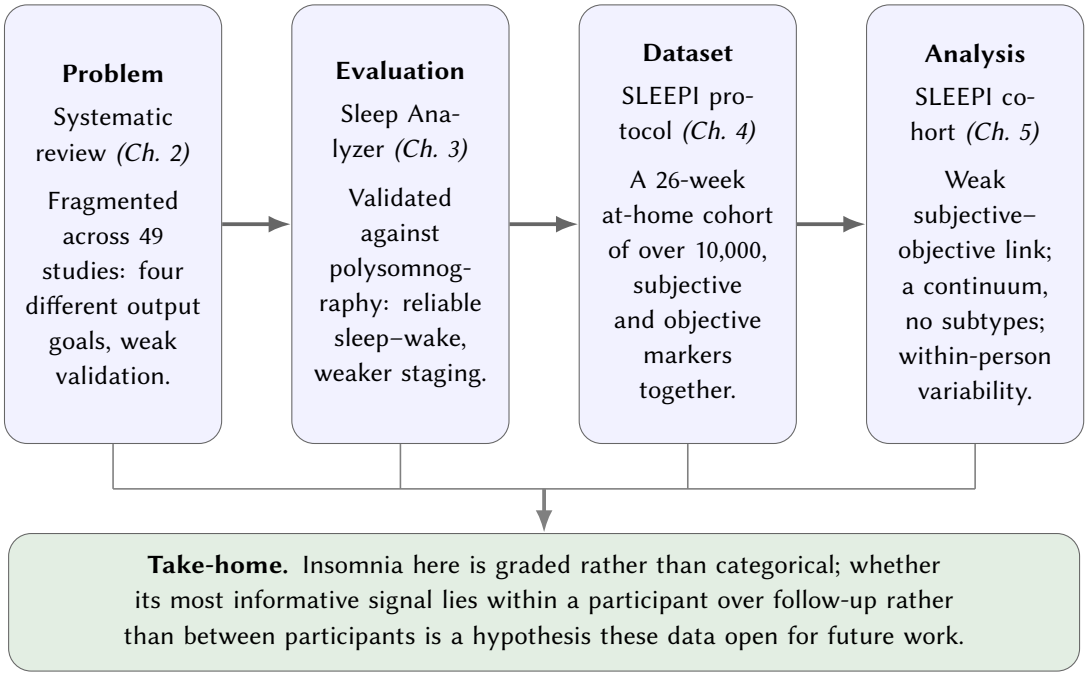


Figure 6.1: Overview of the four contributions and their principal findings. The systematic review documented the fragmented state of digital insomnia assessment and the need for integrated longitudinal measurement, the Sleep Analyzer evaluation quantified how well the objective measurement performs at home, the SLEEPI protocol built the longitudinal dataset recording subjective and objective markers together, and the SLEEPI analyses characterized the insomnia phenotype. Read together, they place insomnia on a severity continuum and open within-person variability over time as a direction for future work.

How insomnia is experienced and how it is measured agree only loosely. Objective insomnia assessed by PSG is more prevalent than its questionnaire-defined counterpart [21], multi-night home actigraphy separates distinct profiles of sleep misperception [110], and the discrepancy between perceived and recorded sleep is recognized across sleep and psychiatric conditions [126]. The literature has centred this divergence on sleep duration: insomnia with objective short sleep duration is held to be the most biologically severe form of the disorder [16], and it is the objective short-sleep phenotype, not its subjective counterpart, that carries about twice the odds of incident hypertension [17]. In the SLEEPI cohort the objective signal that tracks self-reported insomnia severity is weak, and it lies in the fragmentation and circadian markers rather than in sleep duration. Estimating these markers, the circadian ones especially, requires the multi-night longitudinal recording that SLEEPI provides, and the dataset paves the way for further analysis of them and of their change over time. The objective and subjective views thus describe related but complementary constructs.

The design's original contribution is to a dimension a cross-sectional view cannot reach: how a participant changes from night to night and over the weeks of follow-up. Night-to-night variability predicts the response to CBT-I [120], and the variability of sleep around its own mean, long overshadowed by its average, is only recently treated as a target with its own correlates [127]. SLEEPI measures it directly over 26 weeks. The ISI is trait-dominant at an intraclass correlation of 0.75, so about a quarter of its variance is within-person variation rather than stable difference between participants, while the objective features fluctuate from night to night, the standard deviation of total sleep time reaching 1.12 hours. Severity is persistent without being fixed: from one biweekly assessment to the next a participant holds the same tier with probability 0.64 on average, almost all movement confined to adjacent tiers. This within-person axis, accessible only to repeated longitudinal measurement, is where we see the most promising direction for the future characterization of insomnia.

Where variability concerns change within a participant, the question of subgroups concerns structure across participants, and here the objective features show no natural divisions. Stable insomnia subtypes have been derived from non-sleep traits [128], and the heterogeneity of insomnia invites data-driven subtyping [129], yet a latent-class analysis of the ISI recovers a graded continuum from no insomnia to high risk rather than separate kinds [130]. The SLEEPI features behave the same way: no density gap separates them under either standard or non-linear clustering, the low-dimensional projections form a single diffuse cloud, and the best non-degenerate partition agrees with the clinical labels no better than chance, its adjusted Rand index near zero. Where the features do vary smoothly, insomnia severity follows those gradients only weakly, at a Pearson correlation of at most 0.15 with the principal axes of variation, so the dominant structure in the objective features runs largely orthogonal to the insomnia labels. The result bounds its own scope, ruling out discrete subtypes in these per-subject aggregate features rather than a pragmatic stratification by severity or risk, whose boundaries would be a matter of convenience rather than natural division. A supplementary clustering on night-to-night variability rather than on the per-subject medians likewise found no discrete subgroups, so the continuum is not an artifact of describing participants by their typical level; the stable-against-variable distinction that a single-night recording could never reveal proves as graded as the rest [110]. A split on severity would re-describe the ISI rather than reveal an objective subtype, and be-

cause the dominant feature gradients run almost orthogonal to the labels, a grouping that followed them would order participants by something other than insomnia; so if individuals are to be grouped at all, the usable axis is severity or risk rather than a search for objective kinds. Read together with the variability finding, the structure these data carry is temporal, within participants over time, more than typological, between one participant and another.

6.2 Limitations

Several limitations bound the conclusions above. The Sleep Analyzer was validated in healthy adults, so its accuracy in clinical insomnia, and with it the strength of the objective measurement the analyses rest on, remains to be established, particularly because the objective features, and the fragmentation markers most of all, are measured least accurately in disturbed sleep. The cohort was recruited from existing device owners, which selects a particular population and limits how far its demographic and clinical patterns generalize. The insomnia diagnosis label applies the DSM-5 criteria, requiring a reported symptom type and daytime impairment and reclassifying comorbid cases as secondary insomnia, yet it remains a self-reported approximation: it rests on questionnaire answers rather than a clinical interview, the adequate-opportunity criterion has no survey item and is not enforced, and even the strict primary-chronic prevalence near 29% exceeds the roughly 6% of the general population who meet strict clinical criteria [30], so the label still over-selects and every result stratified by it, including the weak subjective–objective association, is bounded by it.

The objective phenotyping used per-subject medians from a single cross-section, on a cohort using one single firmware of one device, so the absence of discrete subgroups holds for these aggregate features; a night-to-night variability aggregation was also checked and showed none, which points to aggregating less, over finer time windows and with better-chosen aggregation rules, as the informative next step rather than a further per-subject summary.

6.3 Perspectives

Several directions follow from this work. With a non-linear autoencoder representation already tested and the feature space found continuous, further unsupervised clustering is unlikely to help; the more promising extensions add supervision or change the input. A supervised characterization of the severity axis would relate the device-derived markers to the validated subjective scales, and a shift from per-subject aggregates to the raw temporal sequences would test whether the aggregation rather than the absence of subgroups is what leaves no separable signal. Because the dominant variation between participants runs almost orthogonal to the insomnia labels, referencing each participant against their own baseline rather than against the cohort would remove the stable between-person differences and test whether the within-person deviations from that personal norm carry the insomnia signal that a cross-participant comparison obscures. We attach particular importance to the within-person dynamics: how severity evolves over follow-up, described by

the tier-to-tier transition probabilities, by individual ISI trajectories, and by whether insomnia persists, remits, or worsens [131, 132]; and the intra-individual variability that may itself serve as a marker [120]. Detection algorithms should be validated in people with insomnia and not in healthy participants alone. The selection bias of a commercial user base should be corrected, and subjective and objective data integrated longitudinally, so that digital biomarkers can be related to established clinical scales and carried toward use in care. The collection design itself, from consent and engagement through dropout handling to decentralized data ingest, is not specific to insomnia; reusable remote-monitoring platforms already span several disorders [133], so the same design could support longitudinal monitoring in other conditions, though only reuse in each new indication would confirm it. The subjective–objective comparison can itself be widened beyond the three summary scores and the complaint-based class, to the survey items left unused, the reported type of sleep difficulty, its daytime consequences and its environmental disruptors, the recorded diagnoses of mental and other sleep disorders, and the individual items of the ISI, PHQ-9 and GAD-7 rather than their total scores. Because sleep apnea was not an exclusion criterion and apnea diagnosis, treatment, and the device-estimated apnea-hypopnea index were recorded, a further direction is to examine how comorbid apnea co-occurs with the insomnia classes and severity tiers in this cohort. The longitudinal design also opens questions about environmental influences on sleep: whether insomnia symptoms follow a seasonal pattern, and how they respond to ambient temperature and weather [134], in particular during heat waves. The diagnosis label already applies the DSM-5 symptom, impairment and exclusion criteria, and could be strengthened further toward a clinical standard, since the adequate-opportunity criterion has no survey item and the classification rests on self-report rather than a clinical interview.

6.4 Conclusion

We combined the validated precision of an at-home sleep-staging evaluation with the scale of a 26-week multimodal collection, and characterized a large ecological cohort of people reporting insomnia symptoms. The contribution speaks to three audiences. For sleep research, we document the fragmentation of digital insomnia assessment and provide a reproducible protocol for building a longitudinal dataset that records subjective and objective markers together. For connected-device research, we quantify how a contactless nearable performs at home against the imperfect reference of manual scoring. For clinical practice, we move toward an individualized, longitudinal, and multimodal characterization of insomnia rather than from a single night. The work is at the research stage, yet it already points in a direction for care: because the objective profile of insomnia forms a continuum rather than sharply bounded classes, a patient may be better placed along a severity gradient than assigned to a discrete phenotype. On a longer horizon, and if the objective markers and their trajectories replicate, the ISI combined with device-derived features could help decide when to refer and, eventually, match a patient to the form of CBT-I that suits their needs [36]. The phenotyping showed no discrete objective subtypes under unsupervised clustering, so insomnia in this cohort reads as a continuum of severity rather than a set of kinds, and the within-person night-to-night variability, which a single cross-sectional

partition cannot represent, is the most original direction the data open. The phenotyping of insomnia from real-world sensor data remains the principal open problem of this work.

Appendix A

Supplementary material for Chapter 2 (systematic review)

A.1 Detailed search queries

Each database has its own query syntax and specificities. The table A.1 summarizes the different platforms used to perform the literature search, the selected, Zotero compatible imported file formats, and the search queries applied to each platform.

Interface Platforms	Imported file format	Search Queries
PubMed (NIH)	PubMed XML	(app OR application OR "machine learning" OR "deep learning" OR "supervised learning" OR "unsupervised learning" OR clustering OR "smart device" OR watch OR ring OR mattress OR headband OR nearable OR wearable) AND (insomnia[tiab] OR "sleep quality"[tiab] OR "sleep disturbance"[tiab] OR "sleep wake"[tiab] OR "sleep disorder"[tiab] OR "sleep loss"[tiab] OR "sleep privation"[tiab] OR "sleep deprivation"[tiab] OR "sleep disruption"[tiab] OR "sleeplessness"[tiab] OR "sleep insufficiency"[tiab]) AND (polysomnography OR photoplethysmography OR pressure OR audio OR sound OR microphone OR actigraphy OR acceler* OR ecg OR electrocardiogram OR ekg OR eog OR electrooculogram OR emg OR electromyogram OR ballistocardiography OR bcg OR piezo* OR eeg OR electroencephalogram)
Web of Science (Clarivate)	BibTex	((((((((((((ALL=(app) OR ALL=(application)) OR ALL=("machine learning")) OR ALL=("deep learning")) OR ALL=("supervised learning")) OR ALL=("unsupervised learning")) OR ALL=(clustering)) OR ALL=("smart device")) OR ALL=(watch)) OR ALL=(ring)) OR ALL=(mattress)) OR ALL=(headband)) OR ALL=(nearable)) OR ALL=(wearable)) AND ((((((((((AB=(insomnia)) OR AB=("sleep quality")) OR AB=("sleep disturbance")) OR AB=("sleep wake")) OR AB=("sleep disorder")) OR AB=("sleep loss")) OR AB=("sleep privation")) OR AB=("sleep deprivation")) OR AB=("sleep disruption")) OR AB=("sleeplessness")) OR AB=("sleep insufficiency")) AND ((((((((((((((((((ALL=(polysomnography)) OR ALL=(photoplethysmography)) OR ALL=(pressure)) OR ALL=(audio)) OR ALL=(sound)) OR ALL=(microphone)) OR ALL=(actigraphy)) OR ALL=(acceler*)) OR ALL=(ecg)) OR ALL=(electrocardiogram)) OR ALL=(ekg)) OR ALL=(emg)) OR ALL=(electromyogram)) OR ALL=(eog)) OR ALL=(electrooculogram)) OR ALL=(piezo*)) OR ALL=(bcg)) OR ALL=(ballistocardiography)) OR ALL=(eeg)) OR ALL=(electroencephalogram))
Scopus (Elsevier)	RIS (Research Information Systems)	(TITLE-ABS-KEY(app) OR TITLE-ABS-KEY(application) OR TITLE-ABS-KEY("machine learning") OR TITLE-ABS-KEY("deep learning") OR TITLE-ABS-KEY("supervised learning") OR TITLE-ABS-KEY("unsupervised learning") OR TITLE-ABS-KEY(clustering) OR TITLE-ABS-KEY("smart device") OR TITLE-ABS-KEY(watch) OR TITLE-ABS-KEY(ring) OR TITLE-ABS-KEY(mattress) OR TITLE-ABS-KEY(headband) OR TITLE-ABS-KEY(nearable) OR TITLE-ABS-KEY(wearable)) AND (TITLE-ABS-KEY(insomnia) OR TITLE-ABS-KEY("sleep quality") OR TITLE-ABS-KEY("sleep disturbance") OR TITLE-ABS-KEY("sleep wake") OR TITLE-ABS-KEY("sleep disorder") OR TITLE-ABS-KEY("sleep loss") OR TITLE-ABS-KEY("sleep privation") OR TITLE-ABS-KEY("sleep deprivation") OR TITLE-ABS-KEY("sleep disruption") OR TITLE-ABS-KEY("sleeplessness") OR TITLE-ABS-KEY("sleep insufficiency")) AND (TITLE-ABS-KEY(polysomnography) OR TITLE-ABS-KEY(photoplethysmography) OR TITLE-ABS-KEY(pressure) OR TITLE-ABS-KEY(audio) OR TITLE-ABS-KEY(sound) OR TITLE-ABS-KEY(microphone) OR TITLE-ABS-KEY(actigraphy) OR TITLE-ABS-KEY(acceler*) OR TITLE-ABS-KEY(ecg) OR TITLE-ABS-KEY(ekg) OR TITLE-ABS-KEY(eog) OR TITLE-ABS-KEY(emg) OR TITLE-ABS-KEY(bcg) OR TITLE-ABS-KEY(piezo*) OR TITLE-ABS-KEY(eeg))
ACM Digital Library	BibTex	[[All: app] OR [All: application] OR [All: "machine learning"] OR [All: "deep learning"] OR [All: "supervised learning"] OR [All: "unsupervised learning"] OR [All: clustering] OR [All: "smart device"] OR [All: watch] OR [All: ring] OR [All: mattress] OR [All: headband] OR [All: nearable] OR [All: wearable]] AND [[Abstract: insomnia] OR [Abstract: "sleep quality"] OR [Abstract: "sleep disturbance"] OR [Abstract: "sleep wake"] OR [Abstract: "sleep disorder"] OR [Abstract: "sleep loss"] OR [Abstract: "sleep privation"] OR [Abstract: "sleep deprivation"] OR [Abstract: "sleep disruption"] OR [Abstract: "sleeplessness"] OR [Abstract: "sleep insufficiency"]] AND [[All: polysomnography] OR [All: photoplethysmography] OR [All: pressure] OR [All: audio] OR [All: sound] OR [All: microphone] OR [All: actigraphy] OR [All: acceler*] OR [All: ecg] OR [All: ekg] OR [All: emg] OR [All: eog] OR [All: piezo*] OR [All: electrocardiogram] OR [All: electromyogram] OR [All:electrooculogram] OR [All: ballistocardiography] OR [All: electroencephalogram] OR [All: bcg] OR [All: eeg]]

Table A.1: Database platforms, imported file formats, and search queries used for the literature review.

A.2 Records included in the full analysis

The 49 studies included in the full analysis of the systematic review are listed in Table A.2.

Table A.2: The 49 records included in the full analysis of the systematic review.

Ref	Authors	Title	Year	DOI
[115]	Abdullah et al.	Support Vector Machine Classification of EEG Nonlinear Features for Primary Insomnia	2018	10.1007/978-981-10-7554-4_28
[46]	Aellen et al.	Disentangling the complex landscape of sleep-wake disorders with data-driven phenotyping: A study of the Bernese center.	2024	10.1111/ene.16026
[95]	Afshani et al.	Discriminating Paradoxical and Psychophysiological Insomnia Based on Structural and Functional Brain Images: A Preliminary Machine Learning Study	2023	10.3390/brainsci13040672
[103]	Andrillon et al.	Revisiting the value of polysomnographic data in insomnia: more than meets the eye	2020	10.1016/j.sleep.2019.12.002
[59]	Angelova et al.	Automated Method for Detecting Acute Insomnia Using Multi-Night Actigraphy Data	2020	10.1109/ACCESS.2020.2988722
[116]	Angelova, Holloway, Ellis	Complexity analysis of human physiological signals based on case studies	2015	10.1088/1742-6596/597/1/012010
[45]	Bonnet, Arand	Activity, arousal, and the MSLT in patients with insomnia.	2000	10.1093/sleep/23.2.1g
[86]	Cheng, Lech, Wilkinson	Simultaneous Sleep Stage and Sleep Disorder Detection from Multimodal Sensors Using Deep Learning	2023	10.3390/s23073468
[106]	Crawford et al.	Characterization of Patients Who Present With Insomnia: Is There Room for a Symptom Cluster-Based Approach?	2017	10.5664/jcsm.6666
[102]	Edinger et al.	The empirical identification of insomnia subtypes: a cluster analytic approach.	1996	10.1093/sleep/19.5.398
[56]	Fossion et al.	Multiscale adaptive analysis of circadian rhythms and intradaily variability: Application to actigraphy time series in acute insomnia subjects	2017	10.1371/journal.pone.0181762

continued on next page

Table A.2 (continued)

Ref	Authors	Title	Year	DOI
[44]	Ghermezian et al.	A Predictive Model to Diagnose Psychophysiological Insomnia with Daytime Hyperarousal and Nighttime Micro–Macro-Structures Correlation	2023	10.1007/s41782-023-00236-5
[118]	Hamida, Ahmed, Penzel	A Novel Insomnia Identification Method Based on Hjorth Parameters	2015	10.1109/isspit.2015.7394397
[112]	Hanif et al.	Automatic Detection of Chronic Insomnia from Polysomnographic and Clinical Variables Using Machine Learning.	2023	10.1109/EMBC40787.2023.10340587
[109]	Hauri	A cluster analysis of insomnia.	1983	10.1093/sleep/6.4.326
[99]	Kusmakar et al.	A machine learning model for multi-night actigraphic detection of chronic insomnia: development and validation of a pre-screening tool	2021	10.1098/rsos.202264
[92]	Lee et al.	Multitask fMRI and machine learning approach improve prediction of differential brain activity pattern in patients with insomnia disorder.	2021	10.1038/s41598-021-88845-w
[107]	Lee, Gonzalez, Small	My job impacts my sleep: signs and symptoms of insomnia among healthcare workers.	2021	10.2486/indhealth.2020-0191
[108]	Lin et al.	Examining Human-Smartphone Interaction as a Proxy for Circadian Rhythm in Patients With Insomnia: Cross-Sectional Study.	2023	10.2196/48044
[105]	Lu et al.	Connectomic disturbances underlying insomnia disorder and predictors of treatment response	2022	10.3389/fnhum.2022.960350
[91]	Manconi et al.	Measuring the error in sleep estimation in normal subjects and in patients with insomnia.	2010	10.1111/j.1365-2869.2009.00801.x
[89]	McCloskey et al.	Data-driven cluster analysis of insomnia disorder with physiology-based qEEG variables	2019	10.1016/j.knosys.2019.07.034

continued on next page

Table A.2 (continued)

Ref	Authors	Title	Year	DOI
[88]	McCloskey et al.	Insomnia Disorder Detection using EEG Sleep Trajectories	2022	10.1007/978-3-031-05981-0_25
[98]	Miller et al.	Clusters of Insomnia Disorder: An Exploratory Cluster Analysis of Objective Sleep Parameters Reveals Differences in Neurocognitive Functioning, Quantitative EEG, and Heart Rate Variability.	2016	10.5665/sleep.6230
[97]	Morgan et al.	Cortical GABA Levels in Primary Insomnia	2012	10.5665/sleep.1880
[113]	Mysliwiec et al.	Comorbid insomnia and obstructive sleep apnea in military personnel: correlation with polysomnographic variables.	2014	10.7205/MILMED-D-13-00396
[60]	Park et al.	Clustering Insomnia Patterns by Data From Wearable Devices: Algorithm Development and Validation Study	2019	10.2196/14473
[68]	Qu et al.	Single-channel EEG based insomnia detection with domain adaptation	2021	10.1016/j.compbiomed.2021.104989
[87]	Rani et al.	Differentiating acute from chronic insomnia with machine learning from actigraphy time series data.	2022	10.3389/fnetp.2022.1036832
[90]	Rodriguez-Morilla et al.	Application of Machine Learning Methods to Ambulatory Circadian Monitoring (ACM) for Discriminating Sleep and Circadian Disorders.	2019	10.3389/fnins.2019.01318
[53]	Shahin et al.	Deep Learning and Insomnia: Assisting Clinicians With Their Diagnosis	2017	10.1109/JBHI.2017.2650199
[83]	Sharma et al.	Automated identification of sleep disorders using wavelet-based features extracted from electrooculogram and electromyogram signals.	2022	10.1016/j.compbiomed.2022.105224
[82]	Sharma et al.	Automated Identification of Sleep Disorder Types Using Triplet Half-Band Filter and Ensemble Machine Learning Techniques with EEG Signals	2021	10.3390/electronics10131531

continued on next page

Table A.2 (continued)

Ref	Authors	Title	Year	DOI
[85]	Sharma et al.	Sleep disorder identification using wavelet scattering on ECG signals	2024	10.1002/ima.22980
[84]	Sharma et al.	Automated insomnia detection using wavelet scattering network technique with single-channel EEG signals	2023	10.1016/j.engappai.2023.106903
[55]	Sharma, Dhiman, Acharya	Automatic identification of insomnia using optimal antisymmetric biorthogonal wavelet filter bank with ECG signals	2021	10.1016/j.compbiomed.2021.104246
[81]	Sharma, Patel, Acharya	Automated identification of insomnia using optimal biorthogonal wavelet transform technique with single-channel EEG signals	2021	10.1016/j.knosys.2021.107078
[110]	Te Lindert et al.	Actigraphic multi-night home-recorded sleep estimates reveal three types of sleep misperception in Insomnia Disorder and good sleepers.	2020	10.1111/jsr.12937
[117]	Thiesse et al.	Somno-Art Software identifies pathology-induced changes in sleep parameters similarly to polysomnography	2023	10.1371/journal.pone.0291593
[94]	Thomas, Wood, Bianchi	Cardiopulmonary coupling spectrogram as an ambulatory clinical biomarker of sleep stability and quality in health, sleep apnea, and insomnia.	2018	10.1093/sleep/zsx196
[101]	Tonetti et al.	Advantages of single step over step-by-step screening for sleep disorders	2018	10.1080/09291016.2017.1390822
[104]	Tsai et al.	Differentiation Model for Insomnia Disorder and the Respiratory Arousal Threshold Phenotype in Obstructive Sleep Apnea in the Taiwanese Population Based on Oximetry and Anthropometric Features.	2021	10.3390/diagnostics12010050
[80]	Wadichar et al.	A Hierarchical Approach for the Diagnosis of Sleep Disorders Using Convolutional Recurrent Neural Network	2023	10.1109/ACCESS.2023.3330901

continued on next page

Table A.2 (continued)

Ref	Authors	Title	Year	DOI
[114]	Wallace et al.	Variable Selection for Skewed Model-Based Clustering: Application to the Identification of Novel Sleep Phenotypes.	2018	10.1080/01621459.2017.1330202
[93]	Warunlawan et al.	Identification of Crucial Factors in Sleep Quality Using Machine Learning Models and MRMR Feature Selection Technique	2023	10.1109/BME-iCON60347.2023.10322011
[111]	Wasiewicz, Skalski	On insomnia analysis using methods of artificial intelligence	2011	10.1117/12.905590
[67]	Wasiewicz, Skalski, Fornal-Pawlowska	Chronic insomnia cases detection with a help of Athens Insomnia Scale and SF-36 health survey	2011	10.1117/12.905587
[96]	Wei et al.	EEG Microstates Indicate Heightened Somatic Awareness in Insomnia: Toward Objective Assessment of Subjective Mental Content.	2018	10.3389/fpsyt.2018.00395
[100]	Zhang et al.	Quantitative evaluation of short-term resting-state brain networks for primary insomnia diagnosis	2022	10.1016/j.bspc.2022.103498

A.3 Detailed inclusion and exclusion criteria

This section gives the full per-study detail behind the exclusion-domain summary of the Datasets subsection of the main text. The summary table there groups these criteria into six domains.

A.3.1 Common exclusion domains

Protocols frequently exclude any neurological or psychiatric history [95, 44, 96, 92, 60], any known comorbidities [59, 87], organic, metabolic, endocrine, or psychiatric disorders [90], or major medical or neurological illnesses confirmed via examination and laboratory testing [97, 92]. More targeted criteria filter out severe or unstable psychiatric conditions and cognitive impairment [98, 92], somatic disorders [96], unmanaged psychosis, substance abuse, and alcohol disorders [99], depression or psychiatric hospitalization within the prior year [45], and traumatic brain injury or gross structural brain abnormalities [95]. Verification methods include structured clinical interviews conducted by trained psychologists or psychiatrists [92, 44, 95, 96, 100] and quantitative psychometric assessments using the Hospital Anxiety and Depression Scale, the Beck Anxiety Inventory, or the Beck Depression Inventory [96].

Secondary sleep disorders represent another heavily restricted domain, with exclusion parameters differing widely among research designs. Exclusions include any diagnosed sleep or circadian disorder [90, 92], unmanaged sleep disorders [99], sleep disorders other than insomnia [98, 92], secondary insomnia [91], and sleep disorders without further specification [101]. Specific conditions targeted include obstructive sleep apnea [94, 96, 102, 91], periodic limb movement disorder, restless legs syndrome, and narcolepsy [96, 44], daytime sleepiness [102], and habitual snoring [94]. Quantitative thresholds differ across studies: an Apnea-Hypopnea Index greater than ten [45] or five [91], a Respiratory Disturbance Index exceeding ten [103], a periodic limb movement frequency exceeding ten per hour [103] or an arousal index greater than ten [45], and restless legs syndrome defined by standard international criteria [91]. Additional exclusions address sleep-disordered breathing, sleep paralysis, nocturia, enuresis, REM behavior disorder, chronic pain, and gastroesophageal reflux disease [97, 104], with visual inspection of polysomnography recordings using American Academy of Sleep Medicine guidelines used as an independent method for identifying comorbid sleep disorders [103].

Pharmacological and substance restrictions isolate primary insomnia from chemically induced sleep disturbances. Protocols prohibit psychiatric [95], neuropsychiatric [44], psychopharmaceutical [101], sedative or hypnotic [91], and psychoactive medications [45], as well as illicit substance or alcohol dependence [98, 105] and high caffeine intake [45]. Screening criteria extend to substance or alcohol abuse within the past month [99] or six months, lifetime substance dependence, or positive urine screening for THC, opiates, cocaine, amphetamines, methamphetamines, benzodiazepines, barbiturates, or PCP [97], alongside restrictions on caffeine exceeding three cups daily or neuroactive prescription medications within the prior three months [97]. Hypnotic-specific criteria exclude indi-

viduals addicted to benzodiazepines [95], those taking sleep medications in the prior two months [96], and those experiencing sleep disruption from the use, abuse, or withdrawal of chemical substances [100]. Washout periods range from two months drug-free [91] to five days for medications influencing sleep or circadian rhythms [92], with some protocols allowing sedative-hypnotics only if discontinued under physician supervision [106] and others forbidding sedative-hypnotics, stimulants, tobacco, marijuana, and more than two alcoholic drinks before bedtime [55]. Completion of a drug or alcohol abuse program within the prior year [45] and secondary insomnia linked to substance abuse [91] also prevent participation.

Physical health exclusions address cardiovascular disease, diabetes mellitus, and respiratory conditions such as asthma and chronic obstructive pulmonary disease [91], unspecified somatic disorders [96, 102], a BMI of 30 kg/m² or greater [94], and pregnancy, lactation, or breastfeeding [99, 98, 105, 95, 92].

Lifestyle, occupational, and circadian variables dictate further restrictions, with studies excluding shift work in the last month [99, 92], flexi-time or night shift work [101], variable, rotating, or split shifts [107], or histories suggestive of circadian desynchrony [45]. To ensure circadian stability, studies exclude subjects demonstrating irregular sleep-wake rhythms verified by one week of actigraphy and concurrent sleep diaries [96] and require a one-week waiting period after transmeridian travel crossing more than one time zone [99]. Exclusions also apply to individuals experiencing continuing domestic abuse [99], those undergoing behavioral treatment therapy in the past month [99], and those receiving concurrent psychotherapy or counseling [105].

Technical and operational parameters further restrict eligibility, excluding participants with contraindications to MRI such as claustrophobia and metallic implants [95, 92] or contraindications for repetitive transcranial magnetic stimulation including implanted metal devices and a history of epilepsy [105], as well as those lacking an Android smartphone [108] or a smartphone capable of running app-based ecological momentary assessment questionnaires [107]. Data quality requirements include sleeping in the laboratory for three consecutive nights [109], completing a sleep diary with at least five of seven consecutive nights [110], providing a one-week continuous actigraphy time series [56], not having left the device unworn for longer than two consecutive days [60], and completing all required questionnaires [109]. Manconi et al. excluded participants whose compromised sleep studies prevented visual sleep stage scoring to avoid misclassification of healthy controls [91].

A.3.2 Insomnia-group-specific criteria

Demographic and environmental parameters exclude individuals older than fifty years [45], bed partners failing to share a bed for over one month [99], individuals whose first-degree relatives are suspected of having depression [97] or whose insomnia is secondary to mental disorders [100], and those receiving ongoing treatment for medical or psychiatric illnesses [60].

Regarding secondary sleep disorders, exclusions target primary complaints of excessive daytime sleepiness or parasomnia coupled with confirmed sleep apnea via polysomnogra-

phy [102], previous surgery for obstructive sleep apnea [104], and excessive sleepiness defined by an Epworth Sleepiness Scale score exceeding sixteen, a score of three on the dozing question, or falling asleep at the wheel within twenty-four months not attributable to acute sleep loss [106]. Clinical thresholds include a Pittsburgh Sleep Quality Index score of seven or higher, an Insomnia Severity Index score of eight or higher [100], insomnia duration shorter than one month, objective total sleep time under one hundred twenty minutes [91], and elevated anxiety or depression on the Hospital Anxiety and Depression Scale, Beck Anxiety Inventory, or Beck Depression Inventory [96].

Participants must be naive to continuous positive airway pressure therapy and CBT-I [106]. One dataset uses a strict occupational inclusion criterion requiring participants to provide direct inpatient care to cancer patients or survivors [107]. Secondary insomnia linked to inadequate sleep hygiene also warrants exclusion [91].

A.4 Detailed subjective-data instruments

This section gives the full per-instrument detail behind the subjective-data summary of the Datasets subsection of the main text, which groups these instruments into categories in Table 2.7.

A.4.1 Public-dataset studies

Subjective sleep metrics collected in these specific investigations include the PSQI questionnaire [92] alongside subjective estimations of standard sleep parameters such as sleep duration, sleep latency (SL), WASO, and the number of awakenings [91]. Researchers additionally track non-standardized psychological metrics encompassing stress, anxiety, and depression levels [93, 92]. Somnolence is quantified using the Epworth Sleepiness Scale [92]. Furthermore, sleep-disruptive cognitions are assessed using the Brief version of the Dysfunctional Beliefs and Attitudes about Sleep Scale-16, which evaluates self-reported faulty appraisals, unrealistic expectations, and perceptual bias regarding sleep [92]. One study additionally collects an unstructured clinical sleep history covering the nature of the complaint (sleep onset, maintenance, or non-restorative sleep), the duration, frequency, and severity of symptoms, daytime functioning, sleep-wake habits, and psychoactive substance intake [90].

A.4.2 Proprietary-dataset studies

Among the studies collecting subjective metrics, researchers use disparate tools, including unstructured clinical histories and specific sleep and psychiatric questionnaires. Standardized diagnostic criteria include DSM-5, ICSD-3, and Research Diagnostic Criteria for Insomnia Disorder parameters [110, 44]. Unstructured evaluations involve psychological, medical, and sleep histories [109], questionnaires about sleep history [102], assessments of sociodemographic characteristics, work characteristics, and perceived stress [107].

Longitudinal tracking relies heavily on sleep diaries spanning daily, seven-day, and two-week periods, including specific formats like the Pittsburgh Sleep Diary [110, 101, 112, 94, 106, 98, 114]. Daily logs document sleep, waking activities, and psychoactive substance use via the Evening-Morning Sleep Questionnaire [97]. Ecological momentary assessment measures capture workday versus non-workday variables including sleep sufficiency, sleep quality, sleep onset latency, nocturnal awakenings, early awakenings, and daytime sleepiness [107]. Other study designs use estimations of home sleep duration and estimated sleep time [113, 46].

Researchers quantify insomnia severity using the Insomnia Severity Index [110, 44, 60, 105] and the Athens Insomnia Scale [67, 111]. Broad sleep quality is evaluated via the Pittsburgh Sleep Quality Index [44, 105, 108]. Sleepiness and fatigue scales incorporate the Epworth Sleepiness Scale [44, 112, 113, 105, 46], the Fatigue Severity Scale [46], and a ten-point visual analog scale quantifying subjective sleepiness [45]. Sleep and wake quality assessments use the mini sleep questionnaire [101].

Psychiatric and mood inventories further expand dataset heterogeneity. Depression scales encompass the Beck Depression Inventory [44], Zung's Depression Inventory [109], and the PHQ-9 [108]. Anxiety is measured via the Beck Anxiety Inventory [44] and Cattell and Scheier's Anxiety Scale [109]. Broad psychiatric evaluations use the Minnesota Multiphasic Personality Inventory, Zuckerman and Lubin's Multiple Affect Adjective Check List [109], and the Profile of Mood States [45]. General health tools encompass the Cornell Medical Inventory [109] and the SF-36 [111].

A.5 Detailed recruitment channels

This section gives the full per-study recruitment-channel detail summarized in the Datasets subsection of the main text.

Among public datasets, participants are primarily sourced from hospital sleep disorder centers, with the eight CAP-based studies drawing from the Ospedale Maggiore of Parma, Italy, and the two MASS-based studies from three hospitals. Individual studies source participants from ongoing epidemiological cohorts [91] or outpatient clinics [92], while the remaining public dataset studies do not indicate their recruitment channels.

Within proprietary datasets, university-affiliated sleep centers and hospital clinics constitute the dominant recruitment pathway [109, 67, 111, 53, 104, 46, 103, 112, 100, 115]. Several protocols adopt divergent recruitment strategies for insomnia and control groups, enrolling patients from hospital clinics while sourcing healthy controls from separate medical centers [94], local advertisements [95], university populations [105], previous research samples [102], or institutional outreach channels [108]. Beyond clinical settings, researchers recruit through online advertisements, radio, print media, and community outreach [99, 44, 110, 45, 98, 97, 96], university boards [60], national sleep registries [96], unspecified databases [101], and ongoing clinical trials [99], while also targeting specific populations such as inpatient nurses [107] and military personnel [113].

A.6 Detailed alternative recording tools

This section gives the full per-device detail behind the alternative-recording-tools summary of the Datasets subsection of the main text.

Rodriguez-Morilla and colleagues record ambulatory circadian data using the Kronowise system, which combines three separate sensors: a Thermochron temperature data logger sampling distal skin temperature every ten minutes, a Hobo accelerometer recording motor activity and body position every 30 seconds, and a Hobo luxometer recording ambient light intensity every 30 seconds [90]. The device-level output is an estimated sleep probability derived from subject-specific thresholds; circadian rhythm indices used as classification features in the study are computed downstream from these raw signals, as described in the Feature Extraction subsection.

Thomas and colleagues use the M1-SleepImage chest wearable, in which one adhesive pad sits beneath the device and a thin wire runs across the chest to a second pad, recording continuous ECG at 600 Hz [94]. Internal accelerometers and gyroscopes estimate body position, and vibration sensors detect snoring. A cardiopulmonary coupling analysis applied to the ECG signal produces CPC sleep metrics, a sleep spectrogram, actigraphic total sleep time, actigraphic wake after sleep onset, and transient awakening counts; these outputs are compared across groups using analysis of variance rather than a classification model.

Lin and colleagues use the Rhythm smartphone application, which passively records the timestamps of screen-on and screen-off events, notification arrivals, and the label of the application in use [108]. From these interaction patterns, the study derives six sleep indicators and three circadian rhythm metrics, as detailed in the Feature Extraction subsection. The study was designed as a validation comparison between app-derived and actigraphy-derived measures, with participants wearing a standard actigraph simultaneously over the monitoring period.

A.7 Detailed objective-data modalities and sampling fidelity

This section gives the full per-modality detail behind the signal-modality summary in Table 2.5 of the Datasets subsection of the main text.

Electroencephalography (EEG) is the primary modality across all datasets. Public repositories typically provide standard clinical sampling rates of 100 to 512 Hz, whereas proprietary datasets frequently use higher-fidelity acquisition of 200 to 1000 Hz. This higher bandwidth in proprietary sources extends to electrooculography and electromyography channels, typically sampled at around 250 Hz, which may enable finer-grained spectral analysis such as beta and gamma band extraction that is lost in down-sampled public records.

Actigraphy is the second most prevalent modality, with a temporal-resolution gap between public and proprietary datasets. Public datasets predominantly provide epoch-based

summary statistics at 0.016 to 0.033 Hz, sufficient for standard sleep-wake estimation according to European guidelines [36], whereas proprietary datasets occasionally provide raw accelerometric data up to 50 Hz, enabling deep learning on motion waveforms. This modality is recommended for longitudinal monitoring in naturalistic settings [36]. Actigraphy devices use configurable wake detection thresholds to discriminate sleep from wakefulness based on acceleration magnitude; one study varied device sensitivity by applying four discrete wake thresholds corresponding to activity counts of 0, 20, 40, and 80 per epoch, generating four parallel feature sets from the same recording to investigate threshold robustness in insomnia classification [59].

Cardiorespiratory and autonomic markers show contrasting availability across repository types. Electrocardiography (ECG) appears in both: four public-dataset studies record it, comprising three drawn from the CAP database [55, 86, 85] and one using the Insomnia-100 dataset [89], alongside seven proprietary-dataset studies [109, 45, 112, 94, 113, 97, 117]; proprietary ECG is sampled at 500 to 600 Hz. Pulse oximetry is confined to proprietary datasets [113, 46], consistent with a focus in some proprietary studies on comorbidities such as Comorbid Insomnia and Sleep Apnea (COMISA), whose assessment requires respiratory-signal analysis.

Environmental and thermoregulatory signals are concentrated in, but not exclusive to, public datasets. Ambient light is recorded in a single public-dataset study [90]. Body temperature appears in one public study, which logs distal skin temperature [90], and in one proprietary study as an oral temperature measurement [45]. Body position is recorded in one public [90] and one proprietary [94] study. Core body temperature and light exposure are proxies of sleep-wake regulation and circadian phase. The limited representation of these modalities in proprietary datasets may constitute a gap in the circadian phenotyping of insomnia. Since modern consumer wearables such as the Oura Ring, the Apple Watch Series 8 and later, the Fitbit Sense, and the Withings SW 2 do measure skin temperature or ambient light, their limited representation in proprietary datasets might indicate that the research groups that collected the data did not consider these signals relevant for insomnia detection, which may reflect a gap in the current understanding of insomnia pathophysiology and its relationship with circadian regulation and could be an avenue for future research. Another possible explanation is that the proprietary datasets were collected before the widespread adoption of consumer wearables and therefore did not include these signals, a pattern that may become more common in more recent datasets falling outside the temporal scope of this review.

A.8 Detailed monitoring duration per subject

This section lists the studies in each monitoring-duration bucket of Figure 2.3 of the main text, using the same four buckets and the same records basis (the dual-dataset study [68] is counted in both the public and proprietary tallies).

- Cross-sectional ($\leq$48 hours): 13 public and 17 proprietary records, comprising single-night and up-to-two-night laboratory recordings, daytime EEG sessions, and single-

session imaging. The imaging studies grouped here use functional MRI in one public [92] and two proprietary [105, 95] studies. This design captures a single sleep episode but cannot assess night-to-night variability or the DSM-5 frequency criterion of at least three symptomatic nights per week.

- Short-term (3–13 days): 4 public and 8 proprietary records. The public studies use week-long data [59, 87] or one-week day-and-night monitoring [90, 56]. The proprietary studies span two to seven nights with night-only [110, 99] or day-and-night [97, 114] recordings, together with [109, 45, 101]; the study with asymmetric per-group windows, namely 3–9 nights for individuals with insomnia versus 10–14 for controls [94], is grouped here, its control arm reaching about two weeks.
- Longitudinal ($\geq$2 weeks): 4 proprietary records, necessary for establishing chronicity but rare: two-week protocols [116, 107] and extended monitoring of 4 weeks [108] to 6 weeks [60].
- Unreported: 1 public [93] and 3 proprietary [104, 67, 111] records did not report monitoring duration, which limits reproducibility.

A.9 Night and diurnal data by study

This section lists the studies in each category of the sleep-wake capture summary of the Datasets subsection, classified by whether the data used for prediction or for the labels are recorded at night, during the day, or both. The dual-use study [68] is counted in both the public and the proprietary tallies.

A.9.1 Publicly available datasets

Fourteen studies use only night data [80, 81, 82, 83, 84, 55, 59, 89, 88, 91, 86, 85, 68, 87], two use only diurnal data [92, 93], and two use both [90, 56].

A.9.2 Proprietary datasets

Nineteen studies use only night data [109, 118, 112, 94, 106, 98, 113, 116, 53, 104, 46, 103, 117, 115, 102, 99, 68, 101, 110], four use only diurnal data [96, 100, 44, 105], and nine use both [45, 60, 108, 107, 114, 97, 95, 67, 111].

A.9.3 Classification notes

Three borderline cases follow the rule that a study is placed by the data used for prediction or for its labels. The study combining MRI with PSG [95] is classed as both, because its labels derive from overnight polysomnography while its predictors are daytime imaging.

The lifestyle-based study [93] is classed as diurnal, because its predictors are systolic blood pressure, body mass index, and physical activity level. The two studies using only Athens Insomnia Scale responses [67, 111] are classed as both, as the scale spans nocturnal and daytime symptoms.

A.10 Detailed validation designs

This section gives the full per-study validation detail behind the paradigm summary in Table 2.11 of the Validation-method subsubsection of the main text.

A.10.1 Supervised model validation

Supervised validation estimates out-of-sample error by scoring a trained classifier on cases withheld from fitting. The 26 supervised studies distribute across four sub-families: resampling without a separately held-out test set in 13 studies, nested resampling combining an inner CV loop with an outer test set in 9 studies, a single fixed train/test partition in 3 studies, and cross-dataset validation training on one cohort and testing on a different cohort in 1 study.

A.10.1.1 Resampling

Resampling validates a classifier by repeatedly partitioning the full dataset into training and evaluation folds and averaging performance across folds. Eight studies apply K-fold CV with K between 5 and 10: five classifiers by the Sharma group on EEG [82, 84], ECG [55, 85], or EOG and EMG [83] signals, a multi-class circadian-monitoring classifier [90], a psychophysiological-insomnia classifier adjusting for depression and anxiety covariates [44], and a functional-connectivity classifier on resting-state functional MRI [105]. Four studies apply LOSO CV, in which each fold withholds every record from a single subject and thereby prevents within-subject leakage across folds: a single-channel EEG classifier [81], a wrist-actigraphy chronic-insomnia classifier [99], a 45-fold design on multi-night actigraphy [59], and a multi-task functional MRI pipeline [92]. One neuroimaging study applies LOO CV, motivated by the small multimodal imaging sample [95].

A.10.1.2 Nested resampling with held-out test set

Nested resampling separates hyperparameter tuning from final performance estimation by combining an inner CV loop on a training partition with a final evaluation on a separately held-out test set. Seven studies apply this design with hyperparameter selection inside the inner loop: a polysomnographic insomnia classifier with 5-fold CV tuning logistic-regression, feedforward-network, and SVM components [112]; a deep-neural-network insomnia detector with hyperparameter tuning on a fixed 8-subject validation set followed

by 5-fold CV repeated five times on the remaining 75 subjects [53]; a three-class classifier distinguishing insomnia disorder from two obstructive-sleep-apnea phenotypes, with 10-fold CV on 80% of the data and a 20% held-out test set [104]; a three-class sleep-quality classifier with 5-fold CV and a held-out validation subset on tabular features [93]; a conditional deep convolutional generative adversarial network insomnia detector with a 75/25 train/test split, 10% of the training partition reserved for validation, and subject-level partitioning [88]; a random-forest polysomnographic insomnia classifier with 3-fold CV repeated ten times on an 80/20 train/test split [103]; and an actigraphy-based chronic-versus-acute insomnia classifier trained on four binary contrasts with 5-fold CV hyperparameter selection on the training partition and a separately held-out test set, repeated five times to average out random sampling [87]. Two further studies use the nested design but are flagged in the extraction for data-leakage concerns: a hierarchical convolutional-recurrent classifier with train/validation/test partitioning and 10-fold CV on a small sample [80], and a multi-modal sleep-stage and sleep-disorder classifier with 90/10 train/test splits and 3-fold CV where training and testing data share subjects [86].

A.10.1.3 Single train/test partition

Three studies evaluate the classifier on a single fixed train/test partition without an additional resampling loop. A single-channel EEG classifier trains on records from ten primary-insomnia and ten control subjects and tests on a separate set of eight subjects per class, thereby partitioning at the subject level [118]. An SVM classifier on non-linear EEG features splits 7500 samples per subject evenly between training and testing halves, with the subject-level partitioning left unstated [115]. A deep-learning classifier on resting-state EEG holds out 50% of samples per subject as test set, which shares subjects across training and testing and thereby admits within-subject leakage [100].

A.10.1.4 Cross-dataset validation

One study validates the classifier's generalizability by training on one cohort and testing on a separate cohort: a domain-adapted long-short-term-memory classifier uses the MASS dataset as source domain containing only healthy subjects and two further datasets with both healthy and insomnia subjects as target domains, with 4-fold LOSO CV applied within each target set [68]. No other included study reports a cross-dataset evaluation.

A.10.2 Unsupervised cluster validation

Unsupervised validation asks whether recovered clusters reflect structure beyond the training sample, either by comparison with independent labels or by model-selection indices computed on the clustering solution itself. The 13 clustering studies distribute across three sub-families: external cluster validity against independent labels in 6 studies, internal indices used solely to select the number of clusters in 6 studies, and one clustering study without a reported validation step. The six studies relying on cluster-count selection crite-

ria do not validate the resulting partition; their methods are described in the Models and outputs subsubsection.

A.10.2.1 External cluster validity against independent labels

Six studies test cluster solutions against external diagnostic categories, phenotypic labels, or prior subtyping drawn from the clinical literature. Hauri interprets nine clusters against external clinical subtype boundaries [109]. Miller and colleagues test three-cluster solutions against external measures including neurocognitive performance, quantitative EEG, and HRV during sleep onset [98]. Mysliwiec and colleagues evaluate three clusters derived from polysomnographic variables against external insomnia-diagnosis categories [113]. Wasiewicz and colleagues compare four Athens-Insomnia-Scale-derived clusters through analysis-of-variance testing on SF-36 quality-of-life profiles as an external reference [111]. Te Lindert and colleagues compare the final latent-class model with insomnia subtypes derived from personality traits and personal history reported in the clinical literature [110]. McCloskey and colleagues report both internal and external validity, the latter through comparison with insomnia-specific EEG phenotypes identified in the same cohort's prior literature [89].

A.10.2.2 No validation reported

One clustering study does not report a validation step beyond the clustering output itself: a clustering study on Athens Insomnia Scale responses combining divisive, fuzzy, hierarchical, agglomerative, k-means, and dendrogram-cutting algorithms [67].

A.10.3 Statistical inference and signal-processing studies

Ten studies fit no predictive model and therefore report no classifier validation; where they test hypotheses, significance is assessed by statistical testing rather than held-out performance. Seven studies state no validation step: a physiological-arousal analysis by analysis of variance [45], a wrist-actigraphy comparison between primary-insomnia and healthy subjects [101], a signal-processing analysis of actigraphy complexity indices [116], a study of smartphone-derived circadian indicators compared with actigraphy through analysis of variance and multi-variate regression [108], a misperception-index analysis combining regression and Bland-Altman plots [91], a signal-processing analysis of actigraphy rhythm and variability [56], and a multilevel-model comparison of nightly sleep characteristics by presence of insomnia-related concerns and by type of day [107]. Two studies record validation as not applicable: a spectroscopy study correlating occipital GABA content with wake after sleep onset [97], and a mixed-model analysis-of-variance comparison between polysomnography and commercial sleep-staging software in which the modeling component is third-party software not evaluated by the authors [117]. One study does not mention a validation step in its reporting: a cardiopulmonary-coupling analysis using analysis of variance with Tukey correction for pairwise comparisons [94].

A.11 Detailed models and outputs

This section gives the full per-study detail behind the output-type summary in Figure 2.5 of the Models-and-outputs subsubsection of the main text.

A.11.1 Binary classification

The targeted insomnia subtype varies: primary insomnia in three studies [115, 118, 101], chronic insomnia in two [99, 105], and psychophysiological insomnia in two [44, 92]. Nine studies target insomnia without specifying a diagnostic subtype [59, 112, 88, 68, 81, 55, 53, 84, 100]. Afshani et al. contrast paradoxical against psychophysiological insomnia without a healthy control group [95].

Supervised machine learning classifiers are used in ten binary classification studies. SVM are the most frequent, appearing in eight studies [115, 95, 59, 44, 112, 99, 92, 55], often alongside alternative classifiers for direct comparison [59, 99]. Random forest is applied in two studies [59, 99], XGBoost in one [105], and ensemble bagged decision trees in one [81]. Hamida et al. apply k-means with two clusters, an unsupervised method categorized here by its binary output structure rather than its learning paradigm [118]. Three studies use deep learning: a conditional deep convolutional generative adversarial network in semi-supervised mode [88], a domain-adapted long short-term memory network [68], and a bidirectional long short-term memory network on functional connectivity features [100]. These three are the deep-learning studies within binary classification, which differ from the three studies applying end-to-end deep learning on raw EEG [88, 68, 80]: the end-to-end set includes the multi-class study [80] but excludes the binary study [100], which operates on extracted connectivity features rather than raw signal. Sharma et al. compare multiple classifiers, with the weighted k-nearest neighbor, ensemble bagged trees, and trilayered neural network yielding the highest performance [84]. Shahin et al. propose a stage-independent deep neural network with subject-level labels derived by majority vote, and a stage-dependent pipeline combining a deep neural network with a hidden Markov model decoded by the Viterbi algorithm; both produce a two-way softmax output [53]. Hanif et al. compare logistic regression, a two-layer feedforward neural network, and an SVM [112]. Tonetti et al. apply fixed clinical thresholds rather than a trained classifier, using the Mini Sleep Questionnaire score and actigraphy-derived sleep onset latency, sleep efficiency, and sleep maintenance arousal to assign subjects to groups [101].

Hanif et al. represent the only study producing probabilistic outputs: the logistic regression and neural network components each yield a continuous probability estimate for insomnia diagnosis, while the SVM produces a hard binary label [112]. Angelova et al. generate predictions at the nightly level across seven nights and derive subject-level labels via a frequency threshold on the proportion of nights classified as bad sleep [59].

A.11.2 Multi-class and multi-task classification

Wadichar et al. first classify subjects as healthy or unhealthy, then reclassify unhealthy subjects into one of five disorder subtypes, yielding six output categories [80]. Sharma and colleagues implement seven-class schemes discriminating healthy status from six sleep disorder types in two studies [82, 85], one of which additionally defines six parallel binary tasks for pairwise disorder discriminability [82]. A further study by the same group produces both binary and multi-class outputs within a single design [83]. Cheng et al. simultaneously output six sleep stage labels and eight pathology class labels, with pathology classes being normal sleep, bruxism, insomnia, narcolepsy, nocturnal frontal lobe epilepsy, periodic limb movement disorder, rapid eye movement (REM) behavior disorder, and sleep-disordered breathing [86]. Rodriguez-Morilla and colleagues discriminate five classes: healthy controls, delayed sleep phase disorder, sleep onset insomnia, maintenance insomnia, and mild insomnia [90]. Warunlawan et al. produce a three-class output distinguishing healthy, insomnia, and apnea [93], and Tsai et al. discriminate insomnia disorder from two obstructive sleep apnea phenotypes defined by low and high arousal threshold [104]. Andrillon et al. apply analysis of covariance for group-effect estimation in addition to random forest classification, addressing three classification contrasts: healthy controls against the combined chronic insomnia group, a three-way contrast among healthy controls and two insomnia subtypes defined by the presence or absence of sleep state misperception, and a binary contrast between those two subtypes [103]. Rani et al. define four binary tasks: acute against chronic insomnia, acute insomnia against the combined control and bed partner group, chronic insomnia against that same group, and pooled insomnia against that group [87]. The inclusion of bed partners in the control group introduces a methodological limitation, as bed partners may themselves exhibit disrupted sleep due to their partner's nocturnal disturbances, which could attenuate separation between the insomnia and control distributions.

Two of the ten studies use deep learning: Wadichar et al. use a convolutional neural network combined with a long short-term memory network in an end-to-end model [80], and Cheng et al. feed six parallel VGG16 convolutional networks into two second-stage shallow neural networks for joint sleep stage and disorder prediction [86]. Warunlawan et al. compare traditional machine learning and deep learning methods, evaluating 21 classifiers including five neural network architectures of increasing depth alongside ensemble, SVM, naive Bayes, and logistic regression models, with ensemble bagged trees achieving the highest accuracy [93]. The remaining seven studies apply traditional supervised machine learning methods: random forest in three [103, 87, 104], with Tsai et al. specifying trees trained by bootstrap sampling [104]; bagged or boosted ensemble decision trees in three studies by Sharma and colleagues [82, 85, 83]; a decision tree trained with the minimum description length criterion in one [90]; and SVM, k-nearest neighbors, and naive Bayes classifiers used comparatively within several of these pipelines [87, 104, 82, 85].

A.11.3 Cluster-based phenotypic stratification

Five studies identify three clusters through distinct algorithms: latent class analysis [110], latent profile analysis [106], variational Bayes Gaussian mixture modeling [89], k-means [113], and hierarchical clustering with Ward's method [98]. The resulting phenotypes reflect different input modalities: Miller et al. separate insomnia subtypes by objective sleep duration and wake patterns [98], Te Lindert et al. identify three sleep misperception profiles [110], Crawford et al. distinguish *High Subjective Wakefulness*, *Mild Insomnia*, and *Insomnia-Related Distress* from diary and questionnaire measures [106], McCloskey et al. define three phenotypes by quantitative EEG spectral features during NREM stage 3 [89], and Mysliwiec et al. separate mixed insomnia with obstructive sleep apnea at two severity levels from predominantly obstructive sleep apnea [113].

k-means is the most prevalent clustering algorithm across included records, appearing in six studies. Hauri identifies nine clusters guided by an empirical interpretability criterion requiring each additional cluster to introduce a clinically distinct subgroup, ranging from psychophysiological insomnia and dysthymia-associated insomnia to childhood-onset insomnia at two severity levels and mapping onto Association of Sleep Disorders Centers nosology categories [109]. Wei et al. identify five EEG microstate clusters corresponding to canonical topographic classes using the Krzanowski-Lai criterion applied at both individual and group level; insomnia subjects show shorter duration in one class and more frequent occurrence in another [96]. Aellen et al. report four clusters for a chronic daily headache cohort and nine clusters each for a sleep-wake disorders cohort and the full combined cohort, imposing a fixed count across cohorts to enable direct comparison; across cohorts, the clusters successfully isolate narcolepsy type 1 and sleep apnea but yield mixed groups for the remaining disorders including insomnia [46]. Park et al. use k-means using synchronic and diachronic temporal strategies, producing six clusters per subject across 35 overlapping eight-day windows; each subject is then assigned to a dominant cluster via PageRank centrality scores, with the resulting clusters capturing distinct sleep-activity profiles such as long sleepers with low sleep efficiency and short sleepers with high efficiency and higher caloric expenditure [60].

Hierarchical clustering with Ward's method is applied in three studies [102, 98, 67]. Edinger et al. derive 14 subgroups of four to 34 subjects each from the reduced feature set, with the optimal solution selected by the cubic clustering criterion and the pseudo-t^2 statistic; the subgroups span phenotypes from normal sleepers with insomnia complaints to insomnia associated with physiological arousal, psychological disturbance, substance use, and periodic limb movements, showing partial concordance with DSM-III-R and ICSD diagnoses [102]. Wasiewicz et al. compare six clustering algorithms, divisive analysis, fuzzy analysis, hierarchical with Ward's method, agglomerative nesting, k-means, and dendrogram cutting, deriving five final groups including one manually constructed group for subjects with the highest difficulty scores, characterizing each cluster by its mean, standard deviation, and the ratio of standard deviations to neighboring cluster mean differences; the five groups range from severe insomnia with high difficulty across all sleep domains to light-level insomnia defined by isolated awakening difficulties [67]. Another study by the same group applies fuzzy analysis clustering to quality-of-life profiles, yielding four clusters labeled severe insomnia, medium insomnia, light evening insomnia, and awaken-

ings insomnia based on SF-36 quality-of-life profiles [111]. Wallace et al. compare mixture models under three distributional assumptions, specifically multivariate normal, multivariate skew-normal, and canonical fundamental skew t, alongside penalized Gaussian mixture models and sparse k-means, with the final solution selected by Bayesian information criterion; the two resulting clusters separate good sleepers with lower SL and more NREM sleep from poor sleepers, with approximately half of the first cluster and three-quarters of the second diagnosed with insomnia [114].

A.11.4 Statistical and signal-level outputs

Among the statistical studies, analysis of variance underlies four approaches: comparison of physiological arousal indices across diagnostic groups [45], three-way analysis of variance with Tukey correction to assess cardiopulmonary measure reproducibility [94], analysis of variance combined with multivariate regression for circadian rhythm and depressive symptom associations [108], and mixed model analysis of variance with Kenward-Roger correction followed by Mann-Whitney U post-hoc tests [117]. The remaining statistical studies use regression-based methods: Morgan et al. characterize the relationship between occipital GABA-to-creatine ratios and polysomnographic wake after sleep onset [97], Manconi et al. fit two normal distributions to misperception index data and apply Bland-Altman analysis with multiple and logistic regression [91], and Lee et al. use multilevel models for continuous and binary sleep variables [107].

The two signal processing studies characterize the temporal and rhythmic structure of actigraphy recordings without classifying subjects. Angelova et al. compute detrended fluctuation analysis, magnitude and sign analysis, power spectral analysis, recurrence quantitative analysis, and variability analysis [116]. Fossion et al. apply cosinor analysis, a redefined intradaily variability index, and singular spectrum analysis to multi-day recordings [56].

A.12 Detailed feature extraction

This section gives the full per-study detail behind the Feature-extraction subsubsection of the main text, complementing Tables 2.9 and 2.10.

A.12.1 Variants of traditional parameters

Beyond the core parameters outlined in Table 2.9, researchers frequently incorporate more specialized, method-specific variations of these macroscopic metrics. Within actigraphy research, additional features capture fine-grained behavioral details, including daily step counts [93], daytime nap durations [107, 46], mid-sleep time [90, 108], and sleep-wake ratio [99, 59, 87]. Other specific actigraphy variables quantify general movement intensity through the sum of all activity, average activity per minute, the standard deviation of activity [114], and a measure of sleep motor activity [101]. Circadian analyses are also expanded

to include the mid-wake time, or MWT, and the integrated TAP variable, which combines temperature, activity, and position [90].

Similarly, PSG recordings frequently yield specific temporal and structural features that go beyond basic sleep architecture. Research works frequently integrate the arousal or wake index [44, 46, 112, 113], sleep maintenance efficiency, or SME, [112], the density of arousals [103], and the frequency of stage transitions [103, 46, 112]. To capture the precise temporal distribution of sleep fragmentation, researchers isolate the minutes awake during the first half, second half, and last two hours of the night [114], alongside the minutes awake specifically after the first occurrence of stage 2 sleep [109]. The structural progression of sleep is further parameterized by calculating the difference between bed time and wake up time, the difference between REM latency and minutes awake prior to REM onset, and the ratio of N3 duration in the first NREM period to N3 duration in the second NREM period [114].

In some studies, specialized parameters are derived to account for confounding clinical factors and atypical sleep phenomena. For example, to quantify the first-night effect, which is a common confounder in laboratory settings, researchers calculate the difference in SL, sleep efficiency, and REM latency between the initial adaptation night and subsequent recordings [109].

Other specific PSG event markers include Spontaneous M1 [44], the exact number of body movements per 100 minutes [109], and the total number of snoring episodes. Furthermore, certain dataset protocols also flag broader sleep-wake abnormalities as discrete categorical features, including the presence of cataplexies, diurnal naps, sleep paralysis, hallucinations, and sleep drunkenness [46].

A.12.2 Mathematical transformations and signal decomposition

In addition to the metrics in Table 2.10, researchers use multi-step or model-based transformations to map latent structures, micro-architectural properties, and functional connectivity. To capture heterogeneous behavioral patterns in commercial actigraphy, unsupervised convolutional autoencoders compress daily features such as sleep timing and caloric expenditure estimates into lower-dimensional representations [60].

Multiple studies used wavelet decomposition to extract frequency subband representations from physiological signals. One record used an optimal filter bank to decompose EEG signals into six subbands [81]. Another study applied a triplet half-band filter pair for wavelet decomposition [82]. One study used a maximally flat multiplier-less biorthogonal filter bank for discrete wavelet transform [83]. Another study applied five-level wavelet decomposition to ECG signals [55].

A methodological discrepancy emerged regarding wavelet transform applicability: one study rejected wavelet analysis on the grounds that it requires a large sample size [118], yet four other studies applied wavelet decomposition to datasets of comparable size without reporting sample-size-related limitations [81, 82, 83, 55]. For EEG analysis targeting delta activity at 0.5 Hz, the 30-second epochs used by these studies provide substantial

margin above the minimum required for wavelet computation; whether the rejection rationale reflects concerns about statistical power for group-level comparisons rather than signal-theoretic constraints cannot be determined from the published text.

One study applied the fast Fourier transform to extract power spectral density values across five frequency bands: delta at 1–4 Hz, theta at 4–8 Hz, alpha at 8–12 Hz, beta at 12–25 Hz, and high-beta at 25–30 Hz [44].

For PSG analysis, wavelet-based methods replace stationary Fourier transforms by providing multi-resolution representations with improved time-frequency localization of transient micro-events, though their application in insomnia classification remains concentrated within a single research group [84, 55, 81, 85]. Continuous time-frequency variations can also be mapped into spectrograms [86], and PCA is applied to manage the high dimensionality of these spectral outputs [89, 102].

Transformations also target specific EEG graphoelements. Researchers isolate slow waves (quantifying frequency, amplitude, and positive slope) and sleep spindles (maximal amplitude, duration, and frequency) to directly measure thalamocortical gating and non-REM sleep depth [103]. Separating these rhythmic oscillations from the background spectrum further enables the estimation of the aperiodic spectrum slope, or $1/f$ noise, which reflects the brain network's underlying excitation-inhibition balance [103].

Eventually, specialized algorithms model brain connectivity and chaotic dynamics. Functional networks are mapped using the Pearson correlation coefficient, Phase Lag Index, Granger Causality or Partial Directed Coherence [100]. The Largest Lyapunov Exponent further quantifies the dynamic unpredictability and underlying chaos of the sleep EEG [115].

A.12.3 Neuroimaging features

Three studies used resting-state functional MRI [105, 95, 92], while one used proton magnetic resonance spectroscopy [97]. One study created individualized parcellations of 180 cortical and 18 subcortical regions from subject-specific atlases derived from diffusion tractography [105]. Another extracted cortical thickness and gray matter volume across 136 anatomical parcels, computing amplitude of low-frequency fluctuations, degree centrality, and regional homogeneity [95]. A third applied region-of-interest PCA on 78 cortical regions, with resulting components processed via LASSO regression and SVM classification [92]. Spectroscopic measures included occipital GABA to creatine ratios and absolute GABA content, integrated with polysomnographic parameters [97].

A.12.4 Subjective instruments

- *Insomnia-specific questionnaires*: The most directly relevant instruments target insomnia severity and sleep-related cognitions. The PSQI has been used to characterize study populations [44]. The ISI, Global Sleep Self-Efficacy Scale (GSES), Basic Adult Sleep Scale (BAS), and Pre-Sleep Arousal Scale (PSAS) were administered in one cohort [106], though notably the ISI served primarily for diagnostic classification into

a priori groups or for purely descriptive statistics in several studies [44, 110, 117, 100], with only Crawford et al. [106] incorporating ISI scores directly as input features for insomnia characterization. The Athens Insomnia Scale (AIS), scored across three symptom dimensions with values ranging from 0 to 3, captured difficulty falling asleep, maintaining sleep during the night, and early morning awakening [67, 111].

- *Psychiatric and mood symptoms*: Broader psychiatric symptoms were quantified using diverse instruments, reflecting recognition that mood disturbance frequently co-occurs with insomnia. Early studies used the Minnesota Multiphasic Personality Inventory (MMPI), Zung's Depression Inventory, Multiple Affect Adjective Check List (MAACL), Anxiety Scale, and Cornell Medical Inventory (CMI) [109]. More recent studies adopted the Beck Depression Inventory (BDI) and Beck Anxiety Inventory (BAI) [44], establishing a temporal measurement of depressive and anxious symptomatology.

- *Daytime symptoms and functioning*: The physiological consequences of poor nocturnal sleep were measured through daytime sleepiness and fatigue scales. The Epworth Sleepiness Scale (ESS) appeared across multiple studies [44, 112, 106, 46], while the Fatigue Severity Scale (FSS) was included in others [106, 46]. The Stanford Sleepiness Scale provided an alternative quantification of daytime somnolence [94].

- *Substance use screening*: Several cohorts incorporated substance use assessment to control for or characterize factors that modulate sleep physiology. In particular, the Alcohol Use Disorders Identification Test (AUDIT) and Fagerstrom Test for Nicotine Dependence (FTND) were administered [105] to quantify alcohol and tobacco use patterns.

- *Subjective sleep parameters*: Direct self-reported sleep metrics complemented objective sleep recordings. Studies extracted subjective total sleep time, including the shortest, mean, and standard deviation of subjective TST [110], or single point-in-time measurements [91, 94]. Sleep diaries spanning seven nights yielded self-reported sleep onset latency, WASO, and sleep efficiency [106]. The same seven-night diary protocol generated means and standard deviations for the number of awakenings, sleep quality ratings, mood upon awakening, and alertness upon awakening [114]. Additional subjective indicators included single-item ratings of sleep quality, physical activity level quantified in daily minutes, and stress level [93].

- *Ecological Momentary Assessment*: A study [107] used ecological momentary assessment to capture real-world variability in subjective experience across 14 consecutive days. The protocol administered four daily prompts with temporally distributed items. The first prompt collected work/free day classification and retrospective assessments of the previous night: sleep sufficiency, sleep quality, sleep onset latency, and the occurrence of nocturnal or early morning awakenings. Three additional prompts assessed daytime sleepiness. This approach captures within-person fluctuations in nocturnal sleep perception and diurnal mood alongside behavioral context.

- *Quality of life*: Broader health-related quality of life was assessed using the SF-36 in at least one cohort [111].

A.13 Detailed preprocessing steps

This section gives the full per-study detail behind the Preprocessing-steps part of the Data-preprocessing subsubsection of the main text.

Signal filtering EEG studies used bandpass filtering to remove low-frequency drift and high-frequency artifacts from muscle and eye movements. Two studies applied a seventh-order Butterworth filter at 0.5–40 Hz [118, 53], one used a finite impulse response filter at 1–35 Hz [82], and one rejected frequencies below 1 Hz and above 30 Hz without specifying filter type [89]. One study applied a 60 Hz notch filter to attenuate powerline interference [44].

Artifact rejection Amplitude-based artifact rejection was used by two EEG studies, which removed epochs containing amplitude values equal to or exceeding 250 μV [118, 53]. For diffusion-weighted neuroimaging, one study eliminated slices exhibiting frame-to-frame intensity changes exceeding two standard deviations from the mean intensity change across all slices [105].

Temporal segmentation Seven studies segmented continuous EEG recordings into fixed-duration epochs of 30 seconds [118, 81, 82, 83, 84, 55, 53], consistent with the AASM scoring conventions . One study additionally applied 50% epoch overlap [118]. Another study removed the initial three minutes of each recording prior to segmentation [82] without stating a rationale for this step. For wearable device data, one study concatenated sleep pattern and daily activity time series across 12 dimensions, combining eight consecutive days of logs as overlapping sliding windows with a stride of one day for subsequent analysis [60].

Normalization and standardization Signal-level standardization was reported in eight studies, though most did not specify the method used. Only one study explicitly stated z-score normalization of the raw signal [55], while six reported standardization without specifying the method [109, 99, 90, 98, 82, 83] and one applied DC-shift elimination to remove baseline drift [118]. Two studies applied z-score normalization to extracted features rather than raw signals [53, 89], and one applied min-max normalization with parameters computed from the training set [87].

Logarithmic transformation One actigraphy study applied $\log_2$ transformation to harmonize recordings from two different actigraphy vendors and reduce skewness effects on feature extraction [87].

Resampling One study subsampled EEG recordings to 256 Hz [84]. One neuroimaging study resliced diffusion-weighted images to ensure isotropic voxels [105].

Discretization One study converted continuous attributes to discrete values using minimum description length discretization, with the authors' rationale that classification models trained on discrete attributes achieve higher accuracy than those trained on continuous attributes [90].

Neuroimaging preprocessing pipeline One functional connectivity study applied a multi-stage preprocessing pipeline for diffusion-weighted imaging, a MRI technique that in-

fers white matter fiber tract orientation from the directional movement of water molecules through neural tissue [105]. The pipeline sequentially corrected head motion via rigid body alignment, removed non-brain tissue, specifically skull and scalp, via convolutional neural network-based skull stripping, corrected spatial distortions from magnetic field inhomogeneities and eddy currents, and estimated fiber orientations via constrained spherical deconvolution. Deterministic tractography then reconstructed structural connectivity pathways from approximately 300,000 streamline samplings, providing the anatomical basis for extracting node parcellations and functional coupling patterns in subsequent analyses.

Appendix B

Supplementary material for Chapter 3 (at-home Sleep Analyzer evaluation study)

B.1 Details on methods and algorithm performance per sleep stage

TABLE OF CONTENTS

Table of Contents
Supplementary Materials A — Methods
Assessment of potential selection bias due to subjects exclusion
Data Processing
Reference
Smoothed, PSG-based Consensus
Synchronization
Definitions: Start and End of the Night
Supplementary Materials B — Performance Evaluation Metrics and Tools
Mean Absolute Error on Sleep Metrics
The mean absolute error (MAE) quantifies the accuracy of a model or prediction by calculating the average of the absolute differences between predicted values and actual values.
Classification Metrics
Plots
Regression
Bland-Altman Plots
Histograms
Boxplots by metadata categories
Supplementary Materials C — Subgroup Analyses
Supplementary Materials D — Results
References

SUPPLEMENTARY MATERIALS A — METHODS

ASSESSMENT OF POTENTIAL SELECTION BIAS DUE TO SUBJECTS EXCLUSION

To assess potential selection bias introduced by data loss, we performed Little's Missing Completely at Random (MCAR) test, comparing the included participants (N = 117) with the excluded group (N = 77). We analyzed four exclusion profiles corresponding to the causes detailed in Figure 1 of the manuscript: (A) missing or incomplete WSA data, (B) missing ECG data, (C) invalid WSA/PSG synchronization, and (D) unreadable EEG signal.

The analysis indicated that exclusions due to reasons A (N = 36), C (N = 11), and D (N = 26) were missing completely at random (p = .23, p = .40, and p = 1.0, respectively). This suggests that data loss coming from WSA technical issues, synchronization errors, or EEG signal quality did not bias the demographic or clinical composition of the final sample. Results are summarized in Table S1.

Exclusion reason	# subjects	Chi-square	Degree of freedom	p-value
A	36	173	162	0.26
B	4	60	41	**2.8e-2**
C	11	54	49	0.29
D	26	132	113	0.11

Table S1: Results of Little's Missing Completely at Random (MCAR) test across exclusion groups
***Bold p-values**: P<.05*

A deviation was observed only for reason B (N = 4, p = 0.04), regarding the mean of PSQI disturbances and medication components.

We used Welch's t-test and chi-squared test to compare metadata mean between the excluded subgroup and the entire population. Compared to the overall study population, subjects excluded for reason B presented lower mean scores for sleep disturbances (1.00 vs 1.27, p = 4e-11) and sleep medication use (0.00 vs 0.20, p = 7e-6), indicating that these excluded participants actually reported slightly better subjective sleep quality and not taking medication. Given the small sample size of this group, this exclusion was not considered a source of systematic bias that would compromise the validation results.

DATA PROCESSING

Reference

Each PSG was read by two trained reviewers following AASM scoring rules [1]. To limit the impact of inter-reader differences [2], epochs with disagreements were adjudicated by a third

qualified professional. The adjudicated review is henceforth called the “reference” or the “consensus”.

Smoothed, PSG-based Consensus

The raw, PSG-based consensus data consists of manually labeled 30-second epochs. WSA outputs a new sleep stage every minute. By design, short stages are smoothed out over 6 minute windows. Therefore, to make a meaningful comparison of the hypnograms produced by WSA and the reference method, we 1) upsampled the output of WSA sleep stage detection algorithm to 30-second epochs by repetition of the value, 2) processed consensus data to get a smoothed version of the PSG-based consensus (see below), and 3) compared the upsampled output of the WSA sleep stage classification algorithm to the smoothed version of the PSG-based consensus. The smoothing process of the PSG-based consensus consists in determining the majority sleep stage over a 6-minute window centered around each 30-second epoch. This approach effectively reduces the variability of the PSG hypnogram by ignoring non-majority epochs within each 6-minute phase, as illustrated on Figure S1. With this method, all events of one sleep stage lasting less than 1min30s and most of those lasting less than 2min30s are removed.

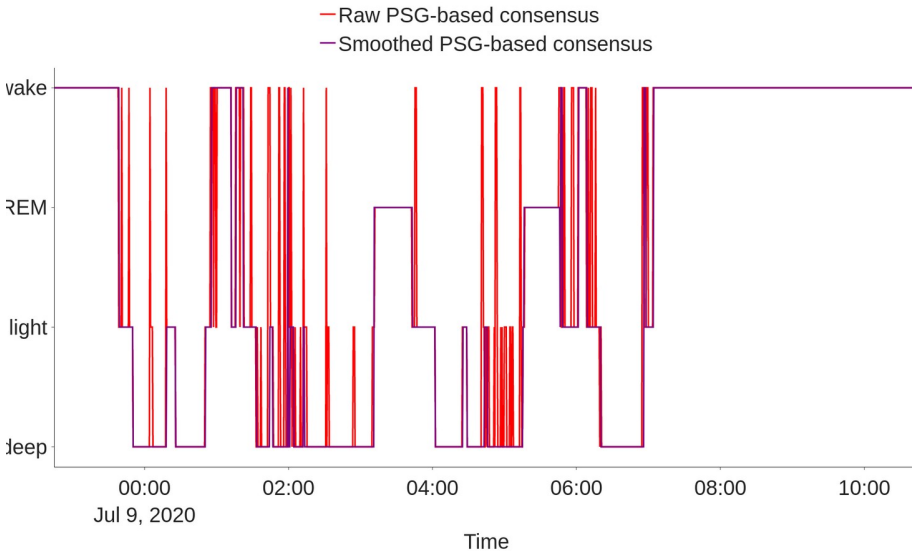


Figure S1: Illustration of consensus hypnogram smoothing (subject 20200708_1)

Synchronization

Recordings from the PSG and the WSA are synchronized using the respiratory signals. The WSA respiratory signal was computed from the normalized pressure signal bandpass-filtered in the frequency range [0.19 Hz, 1 Hz], and the respiratory signal of the PSG was obtained from the ECG filtered similarly [2]. Synchronization then consisted in identifying the optimal affine time warping that maximized their correlation, followed by a visual inspection to ensure that the correlation had a sharp extremum (compare Figures S2 and S3, only the former presents an extremum, for a “shift value” around -5000).

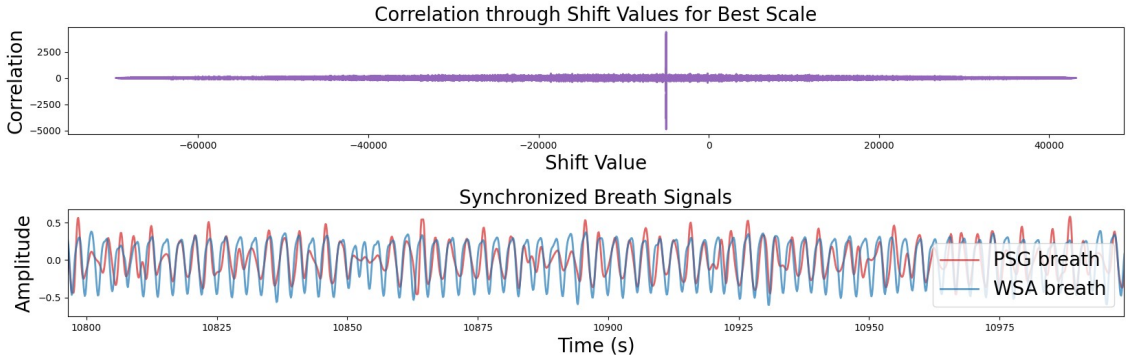


Figure S2: Illustration of a valid synchronization (subject 20201118_1)

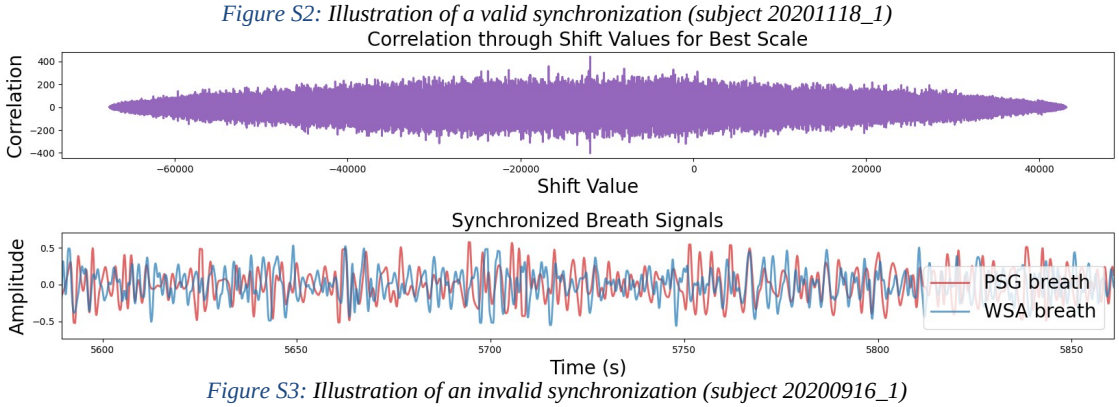


Figure S3: Illustration of an invalid synchronization (subject 20200916_1)

Once determined, these coefficients were applied to the WSA hypnogram data to synchronize it with the reference consensus data.

Definitions: Start and End of the Night

The polysomnograph used in this study recorded 3 EEG, 1 ECG, 2 EOG, 1 submental EMG, and ambient light. In particular, it did not have a position sensor, so we do not have a reference for time in bed and sleep related measures depending on it (bedtime, sleep latency, wakefulness after sleep onset, sleep efficiency, etc.).

In the absence of an indicator of the subject's presence in bed in PSG data, lights-off/on events were used as a proxy to estimate bed presence. If participants went to bed before starting the PSG, or got up after stopping the PSG, the WSA might detect presence without corresponding PSG data. Such unmatched data were excluded from the analysis, relying solely on WSA data that could be compared to the reference.

SUPPLEMENTARY MATERIALS B — PERFORMANCE EVALUATION METRICS AND TOOLS

Mean Absolute Error on Sleep Metrics

The mean absolute error (MAE) quantifies the accuracy of a model or prediction by calculating the average of the absolute differences between predicted values and actual values.

Classification Metrics

We assess the classification of sleep and wake phases, and the classification of sleep stages. Standard classification metrics are used in this work: accuracy, unweighted kappa, sensitivity, specificity. These metrics were calculated for each night individually and then averaged across all nights. These metrics, computed with the ground truth obtained from the PSG and predictions from the WSA, provide insights into the algorithm's ability to correctly classify sleep stages. Below are the definitions and descriptions of each metrics.

Confusion Matrix Example

To illustrate these metrics, consider a binary classification scenario (e.g., sleep vs. wake) as in Table S2, using the following acronyms:

- True Positives (TP): Instances correctly identified as wake
- False Positives (FP): Instances incorrectly identified as wake
- True Negatives (TN): Instances correctly identified as sleep
- False Negatives (FN): Instances incorrectly identified as sleep

		Ground Truth (PSG)	
		Sleep	Wake
Predictions (WSA)	Sleep	TN	FN
	Wake	FP	TP

Table S2: Confusion matrix of a sleep-wake classification

This confusion matrix serves as a valuable tool for evaluating the performance of the classification model, providing a clear visualization of the algorithm's strengths and weaknesses in identifying sleep and wake states.

Sensitivity (Recall) S_e: The proportion of actual positive cases (eg, awake time) correctly identified by the Device Under Test (DUT). It is calculated as:

$$S_e = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Specificity S_p: The proportion of actual negative cases (eg, sleep time) correctly identified by the DUT. It is calculated as:

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

Accuracy: The proportion of all cases (both positive and negative) that are correctly identified. It is calculated as:

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{Total}}$$

Kappa: A statistical measure of inter-rater agreement for qualitative (categorical) items, accounting for agreement occurring by chance.

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where:
P_o is the observed agreement, which is the proportion of instances where the raters agree (ie, the sum of the diagonal elements of the confusion matrix divided by the total number of instances).
P_e is the expected agreement by chance, calculated based on the marginal totals of the confusion matrix.

These metrics were calculated for each night individually and then averaged across all nights. This method captures night-to-night variability, providing a nuanced assessment. It allows for detailed analysis of specific nights, offering granular insights into patterns or issues.

Plots

Regression

Ordinary Least Squares (OLS) regression was used to model the relationship between estimated sleep metrics and reference measurements. The data were assessed for normality and homoscedasticity of the residuals were assessed. Normality was evaluated using the Shapiro-Wilk test, while homoscedasticity was assessed using Levene's test. Both tests confirmed that the assumptions of normality and homoscedasticity were met for all metrics except total sleep time (TST), allowing for the application of parametric statistical methods in subsequent analyses. Regarding TST, RANSAC (Random sample consensus) regression has been performed as it is robust to outliers and does not assume constant error variance, making it suitable for data with varying error distributions [5].

For each regression, the following values are provided:

- R-squared value: “proportion of variance in the outcome variable which is explained by the predictor variables in the sample” [3]
- p-values of the slope and the intercept: “probability under the assumption of no effect or no difference (null hypothesis), of obtaining a result equal to or more extreme than what was actually observed” [4]

Bland-Altman Plots

Bland-Altman plots are used to assess the agreement between two different measurement techniques. They display the difference between the measurements on the y-axis against their mean on the x-axis. Bland-Altman plots are an effective method for comparing estimated sleep metrics to reference measurements by visually assessing the agreement between the two. They highlight the mean difference and the limits of agreement, revealing any systematic bias or variability.

Histograms

Histograms are used to visualize the underlying frequency distribution of a continuous variable. They display the count of observations on the y-axis against defined value ranges bins on the x-axis. In this context, histograms provide a detailed view of the distribution of measurement errors, offering insights beyond summary statistics. They effectively highlight the central tendency and spread of the errors, facilitating the identification of skewness or outliers in the device's performance.

Boxplots by metadata categories

Boxplots summarize the distribution of a dataset by displaying its median, quartiles, and potential outliers. They are useful in this context because they provide a clear visual comparison of the performance across different population subgroups.

SUPPLEMENTARY MATERIALS C — SUBGROUP ANALYSES

Subgroup analyses were conducted to determine whether WSA performance is affected by the following factors: 1) demographic: age, BMI, sex, 2) sleep quality factors: PSQI score, PSQI components, hypnotics, reported sleep quality, and whether the participant has woken up at least once during the night of the recording, 3) environmental: whether the subject has a bed partner, mattress type, mattress thickness. The subgroup categories are summarized in Table S3.

204

Metadata name	Meaning	Categories
age_group	Age category the participant belongs	18-30, 30-50, 50+
bmi_group	BMI category the participant belongs	<18.5, 18.5-25, 25-30, >30
sex	Participant's sex	man, woman
is_couple	Did the participant have a bed partner the night of the recording?	no, yes
psqi_score_group	Results obtained using the Pittsburgh Sleep Quality Index (PSQI) survey, categorized into 4 groups	no_sleep_difficulty, mild_sleep_difficulty, moderate_sleep_difficulty, severe_sleep_difficulty
psqi_quality_group	PSQI Component 1: Subjective sleep quality	good, bad
psqi_latency_group	PSQI Component 2: Sleep latency	good, bad
psqi_duration_group	PSQI Component 3: Sleep duration	good, bad
psqi_efficiency_group	PSQI Component 4: Habitual sleep efficiency	good, bad
psqi_disturbances_group	PSQI Component 5: Sleep disturbances	good, bad
psqi_medication_group	PSQI Component 6: Use of sleeping medications	good, bad
psqi_day_dysfunction_group	PSQI Component 7: Daytime dysfunction	good, bad
hypnotics	Did the participant take hypnotics the night of the recording?	no, yes
sleep_quality	Self-reported sleep quality the night of the recording	very good, good, quite good, bad, very bad
mattress	Type of the mattress the participant slept on the night of the recording	spring, memory foam, foam, latex, other, unknown
mattress_thickness_group	Thickness of the mattress the participant slept on the night of the recording	thin (<20 cm), thick (≥20 cm)
had_arousals	Did the participant wake up at least once during the night of the recording (self-report)?	no, yes

205 *Table S3: Summary of subgroup categories used in boxplots obtained on standard sleep-related metrics*

SUPPLEMENTARY MATERIALS D — RESULTS

Sleep Quality Measures	Slope (p-value)	Intercept (p-value)	R-squared
TST	0.98* (**7e-36**)	0.12* (7e-1)	0.89*
TIREM	0.65 (**7e-8**)	0.36 (**8e-2**)	0.25
TILight	0.41 (**3e-4**)	3.09 (**9e-15**)	1.10e-1
TIDeep	0.02 (8e-1)	1.77 (**1e-7**)	4.00e-4
TIWake	0.73 (**2e-20**)	0.47 (**2e-2**)	0.5
PIREM	0.72 (**5e-8**)	2.33 (3e-1)	0.23
PILight	0.40 (**3e-4**)	0.40 (**5e-16**)	0.11
PIDeep	0.09 (5e-1)	16 (**2e-6**)	5.00e-3
PIWake	0.67 (**4e-15**)	6.25 (**7e-3**)	0.4

Table S4: Summary of regression results obtained on sleep-related metrics between WSA and PSG estimations (N=117)
**Results obtained with RANSAC method. Other regression results are obtained using the Ordinary least squares (OLS) method.*
***Bold p-values**: P<.05*

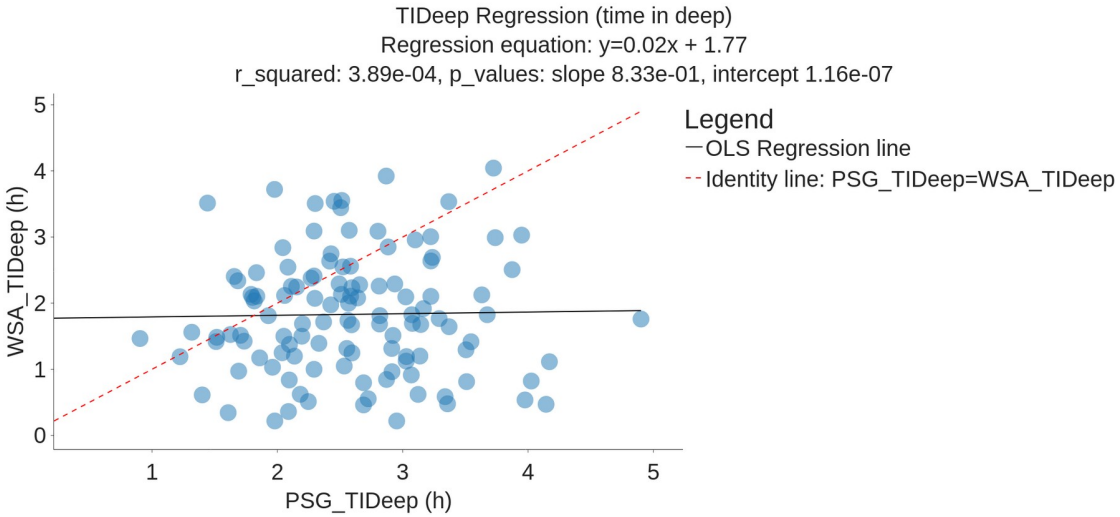


Figure S4: Time In Deep sleep (TIDeep): OLS Regression plot between WSA and PSG (black line) compared to identity line (red dashed line) on 117 subjects (r^2 = 0.02, slope p-value: 8e-1)

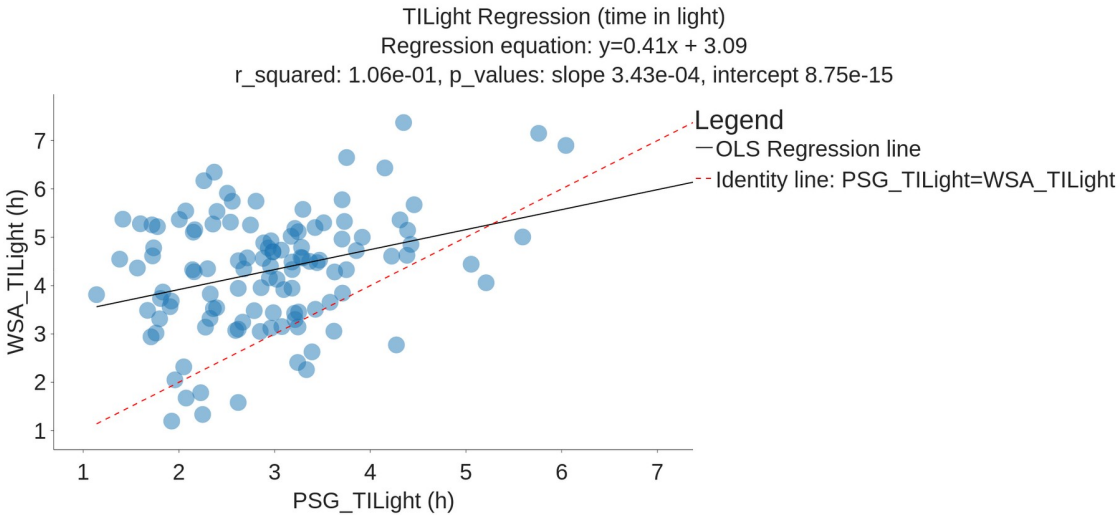


Figure S5: Time In Light sleep (TILight): OLS Regression plot between WSA and PSG (black line) compared to identity line (red dashed line) on 117 subjects (r^2 = 0.41, slope p-value: 3e-4)

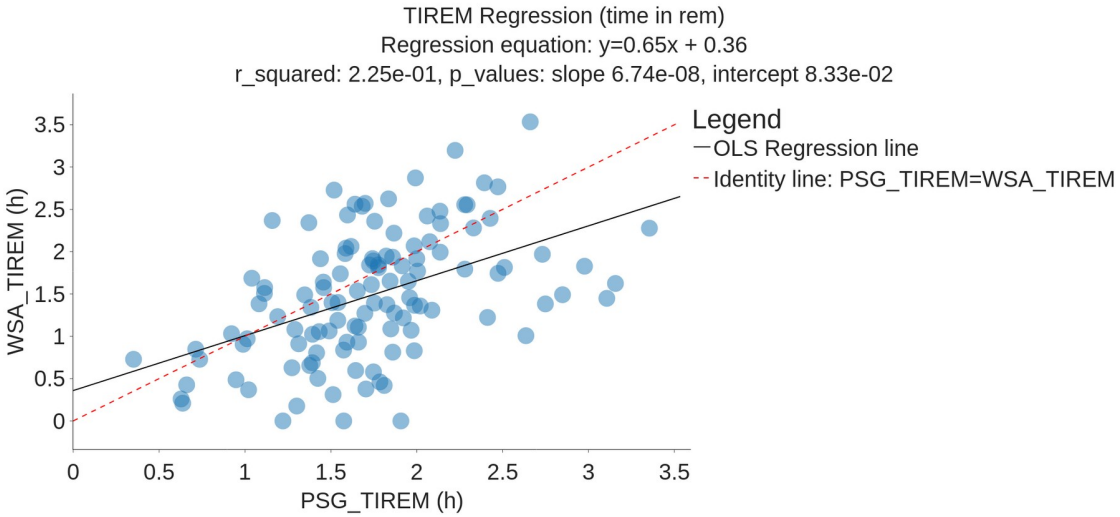


Figure S6: Time In REM (TIREM): OLS Regression plot between WSA and PSG (black line) compared to the identity line (red dashed line) on 117 subjects (r^2 = 0.65, slope p-value: 7e-8)

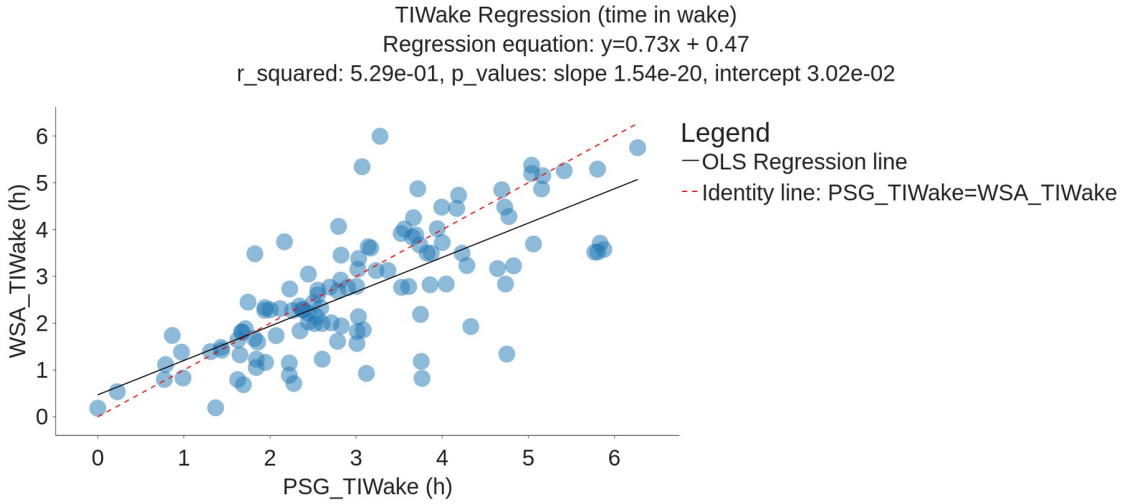


Figure S7: Time In Wake sleep (TIWake): OLS Regression plot between WSA and PSG (black line) compared to identity line (red dashed line) on 117 subjects (r^2 = 0.53, slope p-value: 2e-20)

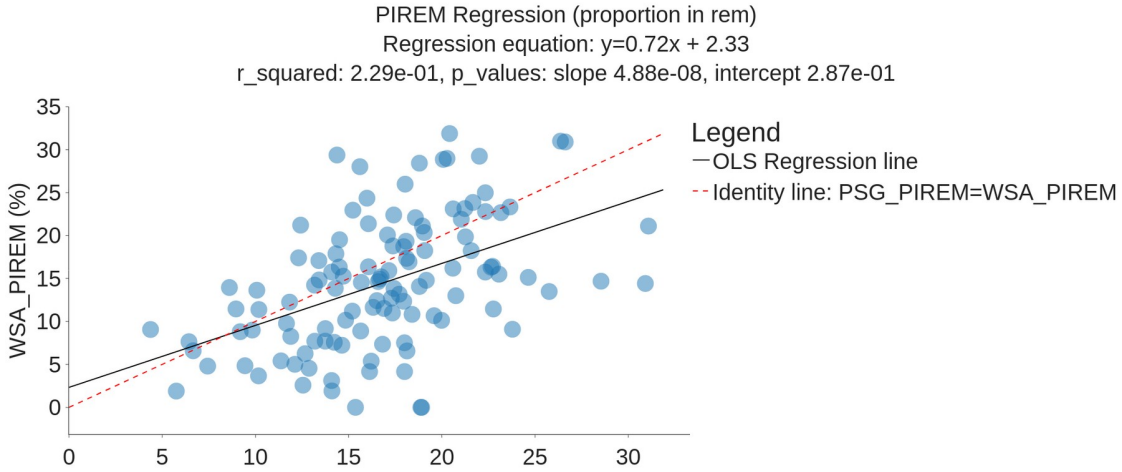


Figure S8: Proportion In REM (PIREM): OLS Regression plot between WSA and PSG (black line) compared to identity line (red dashed line) on 117 subjects (r^2 = 0.23, slope p-value: 5e-8)

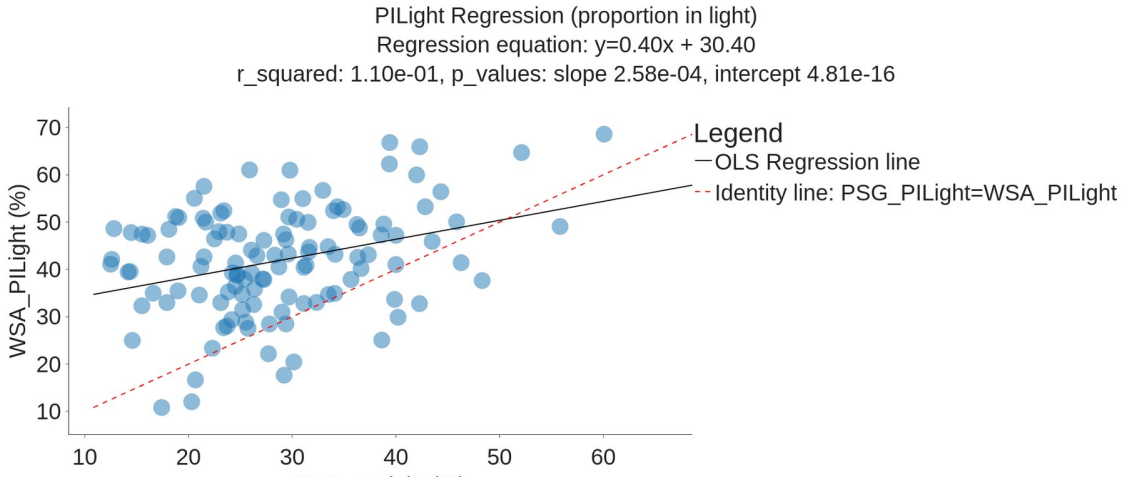


Figure S9: Proportion In Light sleep: OLS Regression plot between WSA and PSG (black line) compared to identity line (red dashed line) on 117 subjects (r^2 = 0.11, slope p-value: 3e-4)

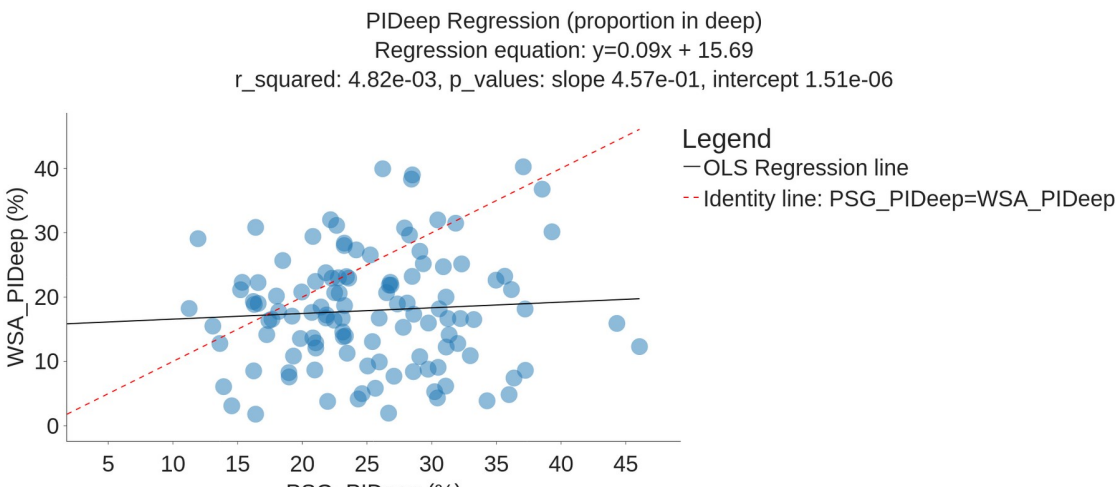


Figure S10: Proportion In Deep sleep (PIDeep): OLS Regression plot between WSA and PSG (black line) compared to identity line (red dashed line) on 117 subjects (r^2 = 5e-3, slope p-value: 5e-1)

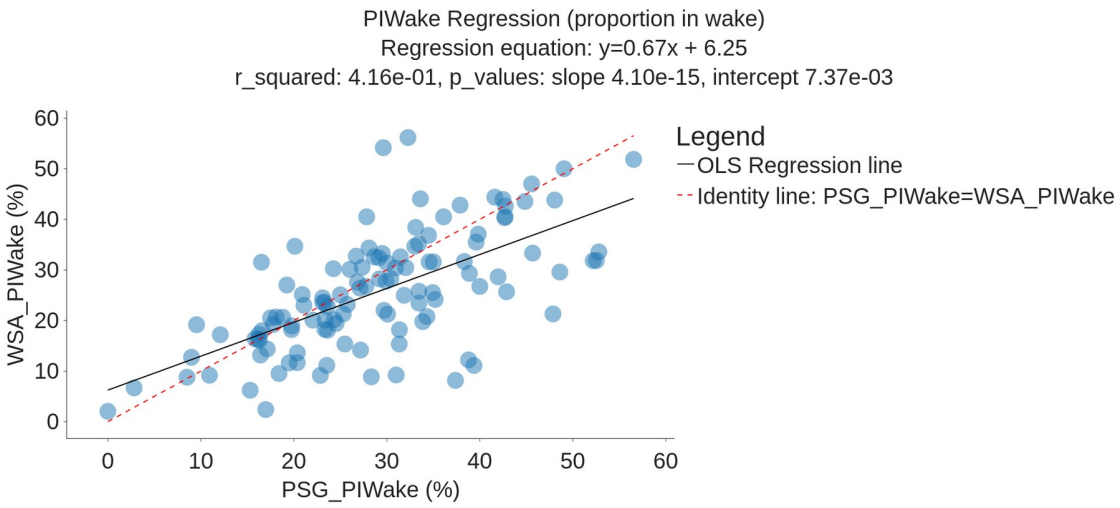


Figure S11: Proportion In Wake sleep (PIWake): OLS Regression plot between WSA and PSG (black line) compared to identity line (red dashed line) on 117 subjects (r^2 = 0.42, slope p-value: 4e-15)

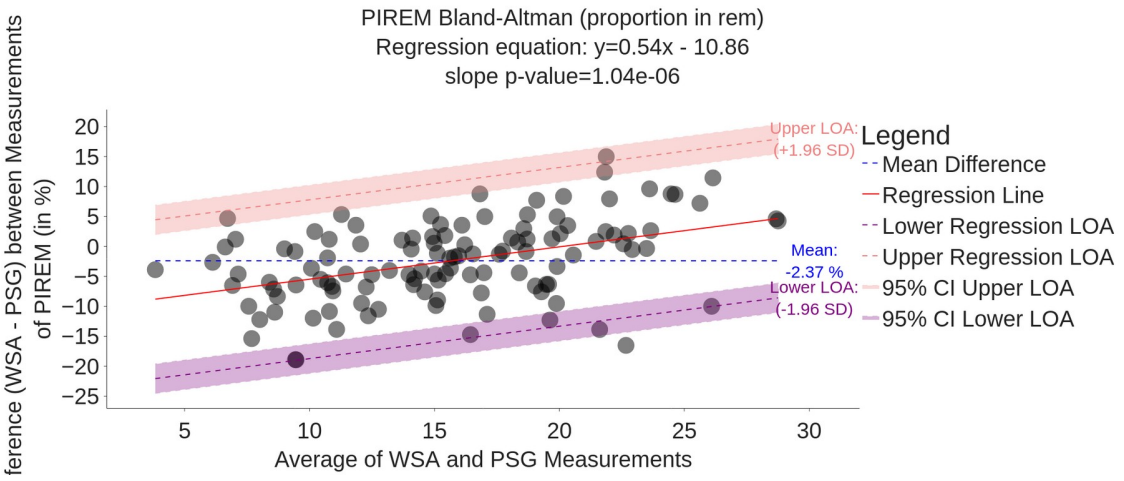


Figure S12: Bland-Altman Plot Comparing WSA and PSG Estimation of PIREM (%) on 117 subjects

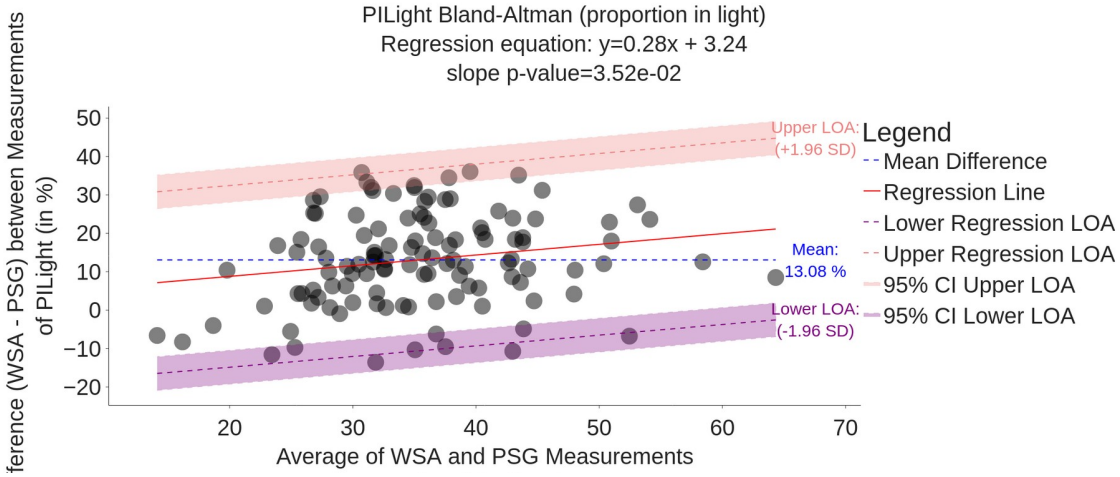


Figure S13: Bland-Altman Plot Comparing WSA and PSG Estimation of PILight (%) on 117 subjects

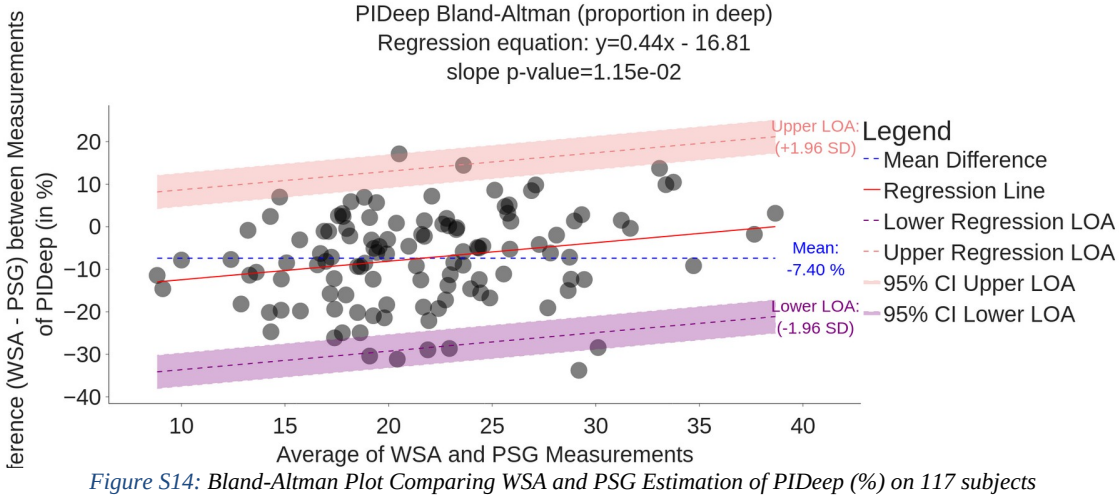


Figure S14: Bland-Altman Plot Comparing WSA and PSG Estimation of PIDeep (%) on 117 subjects

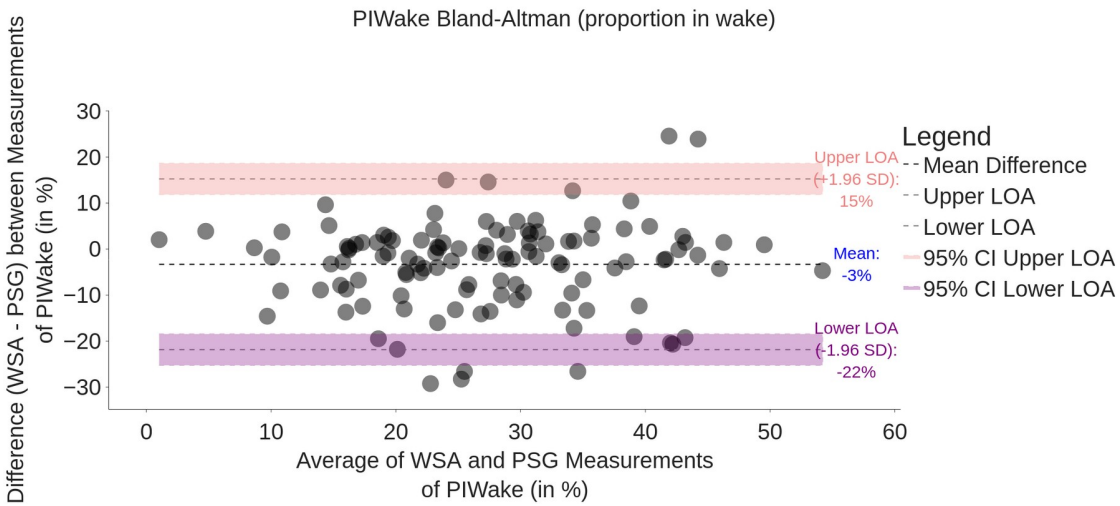


Figure S15: Bland-Altman Plot Comparing WSA and PSG Estimation of PIWake (%) on 117 subjects

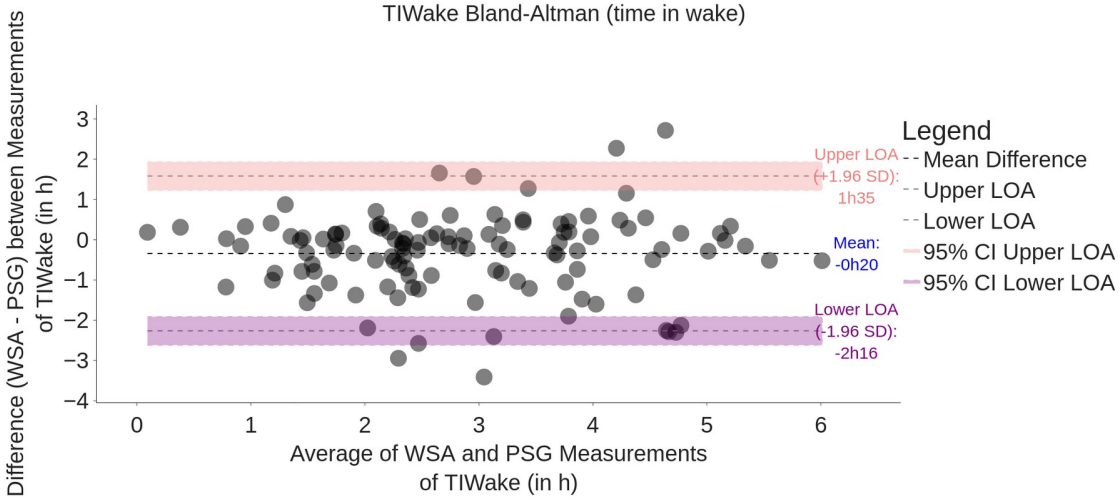


Figure S16: Bland-Altman Plot Comparing WSA and PSG Estimation of time in wake stage on 117 subjects

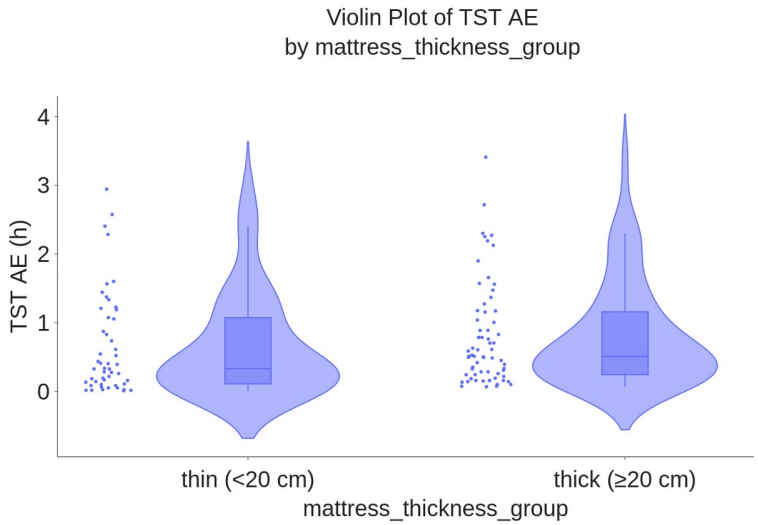


Figure S17: Boxplot of TST AE (h) by mattress thickness group on 117 subjects

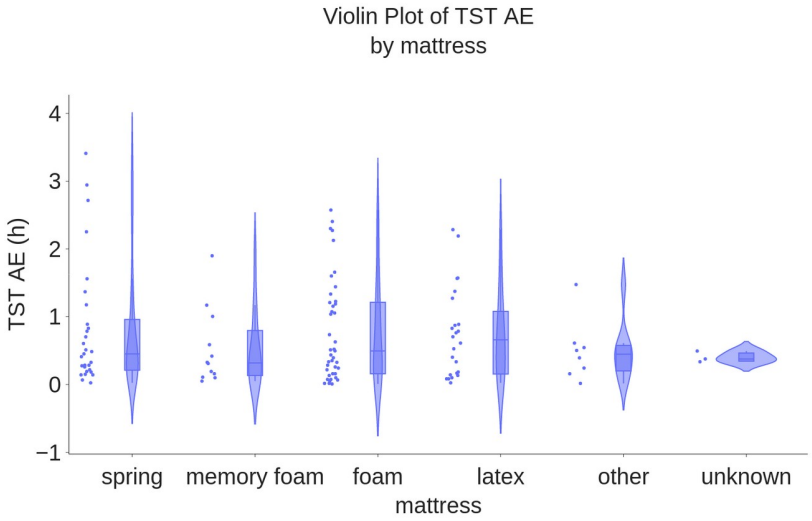


Figure S18: Boxplot of TST AE (h) by mattress type group on 117 subjects

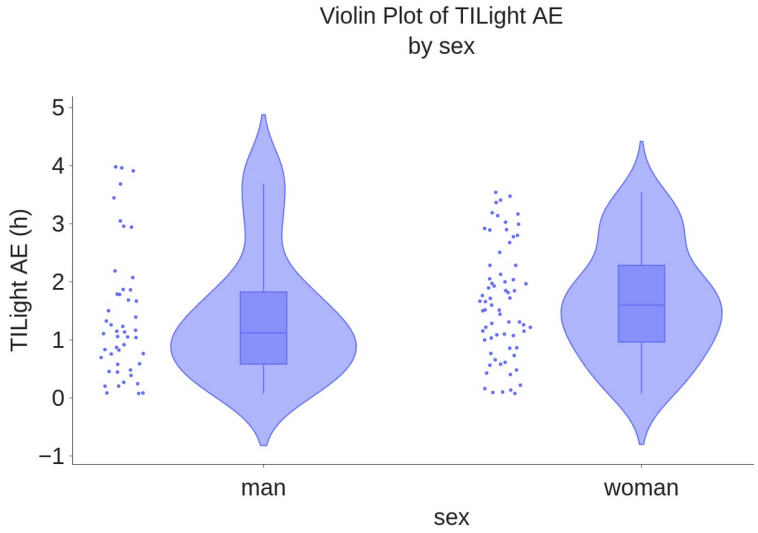


Figure S19: Boxplot of TST AE (h) by sex group on 117 subjects

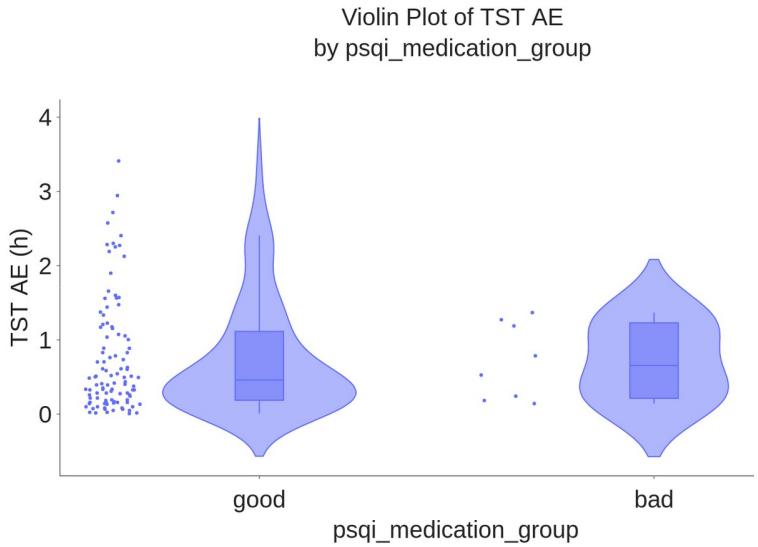


Figure S20: Boxplot of TST AE (h) by PSQI medication group on 117 subjects

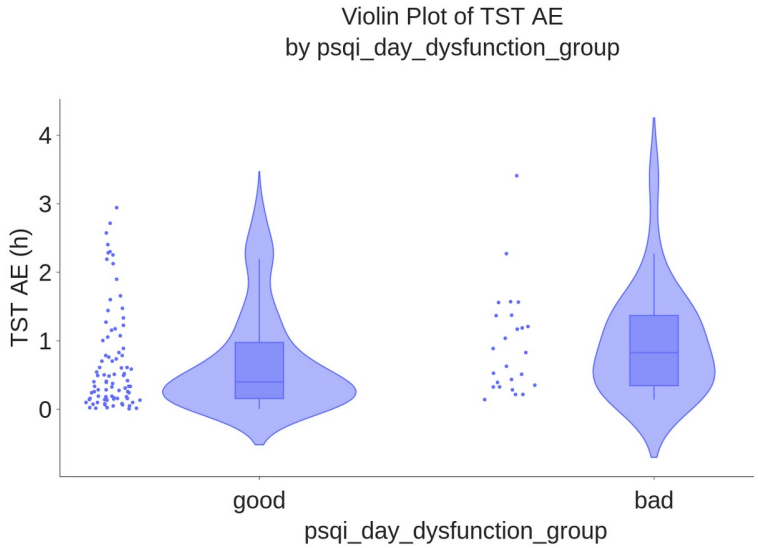


Figure S21: Boxplot of TST AE (h) by PSQI day dysfunction group on 117 subjects

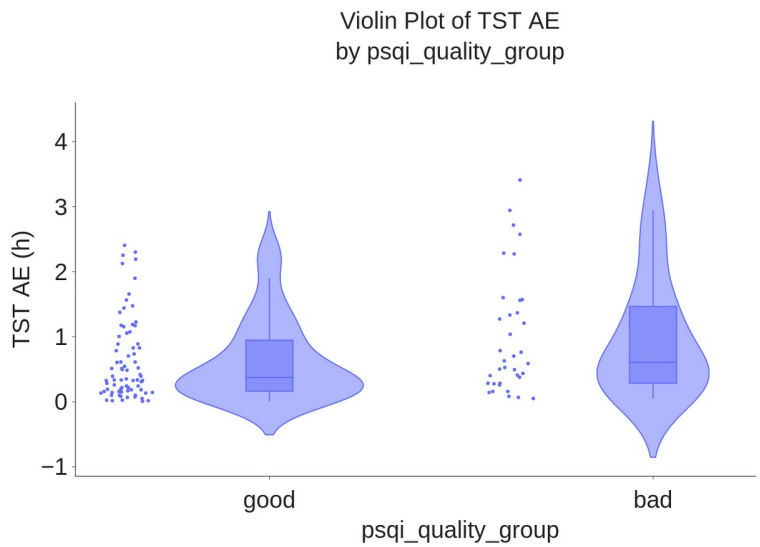


Figure S22: Boxplot of TST AE (h) by PSQI sleep quality group on 117 subjects

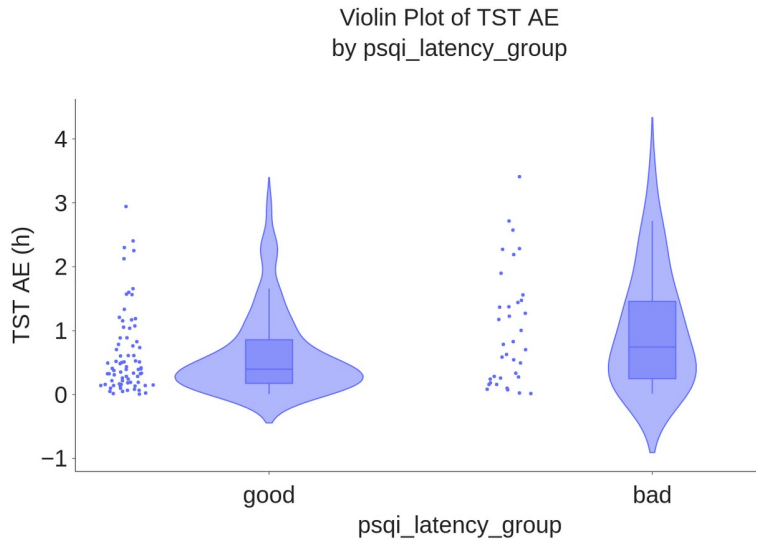


Figure S23: Boxplot of TST AE (h) by PSQI sleep latency group on 117 subjects

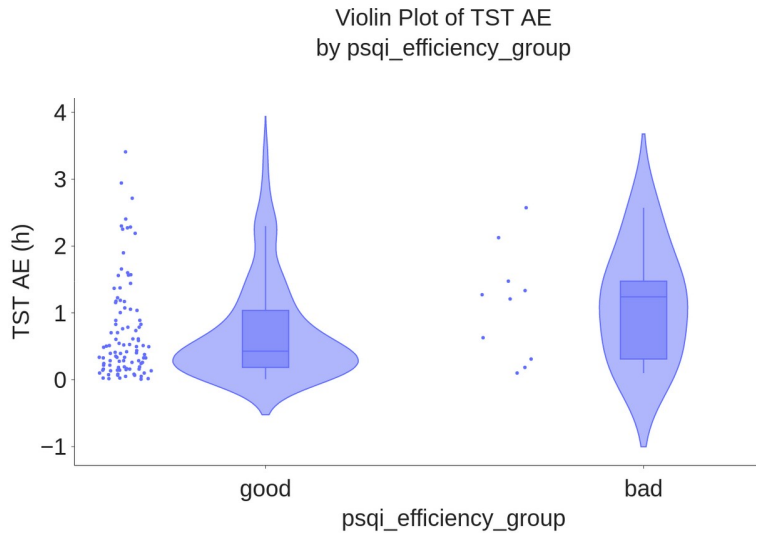


Figure S24: Boxplot of TST AE (h) by PSQI sleep efficiency group on 117 subjects

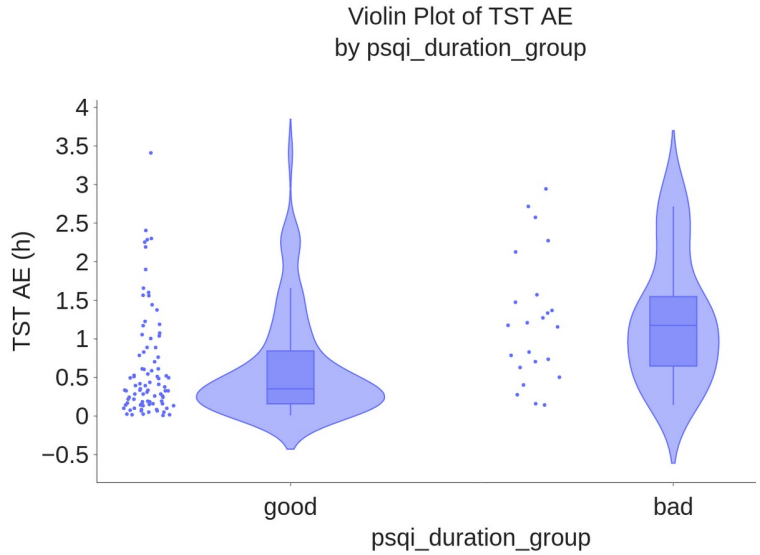


Figure S25: Boxplot of TST AE (h) by PSQI sleep suration group on 117 subjects

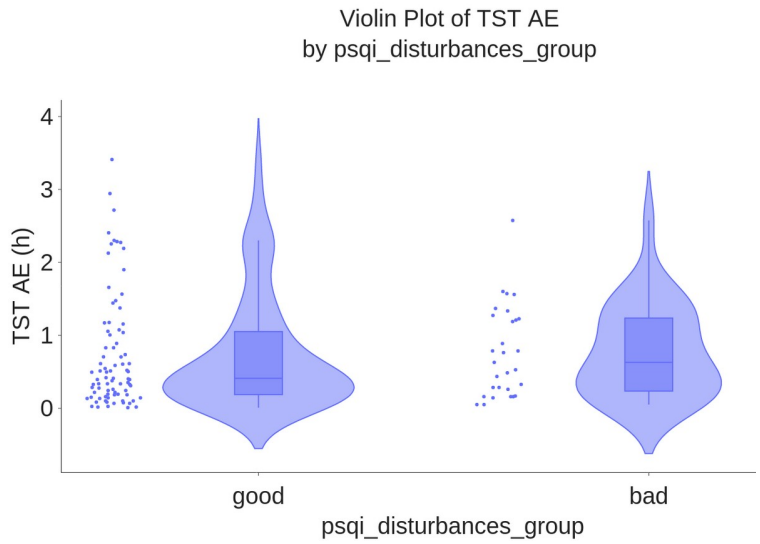


Figure S26: Boxplot of TST AE (h) by PSQI sleep disturbances group on 117 subjects

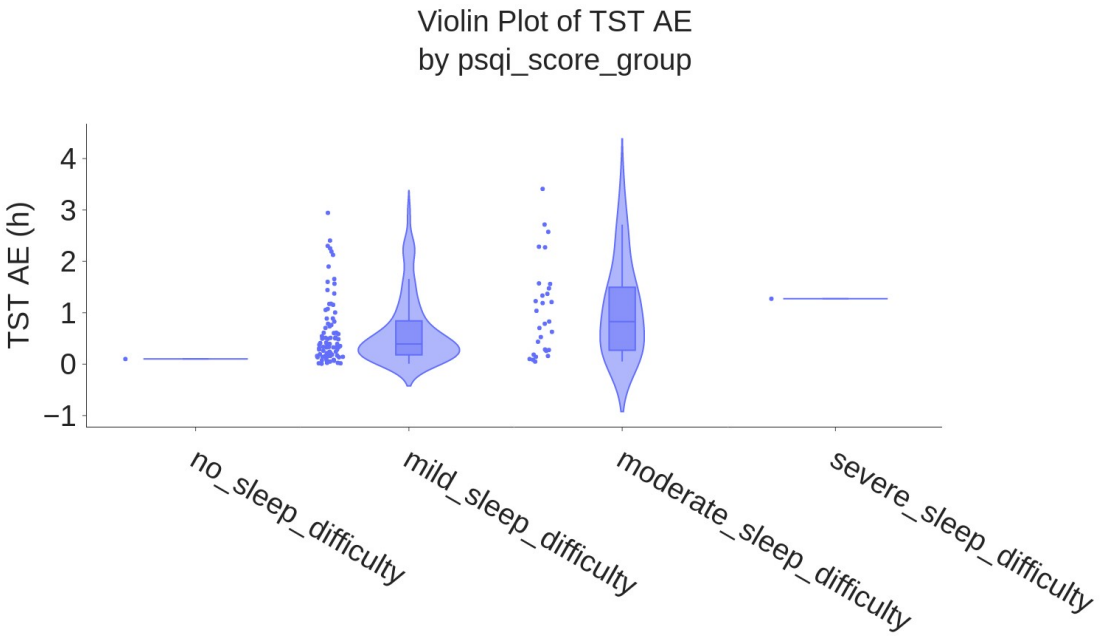


Figure S27: Boxplot of TST AE (h) by PSQI score group on 117 subjects

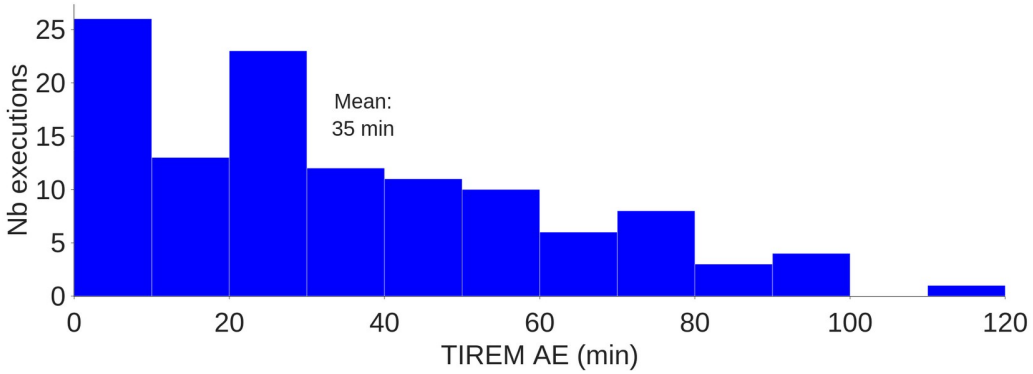


Figure S28: Histogram of TIREM Absolute Error (min) on 117 subjects

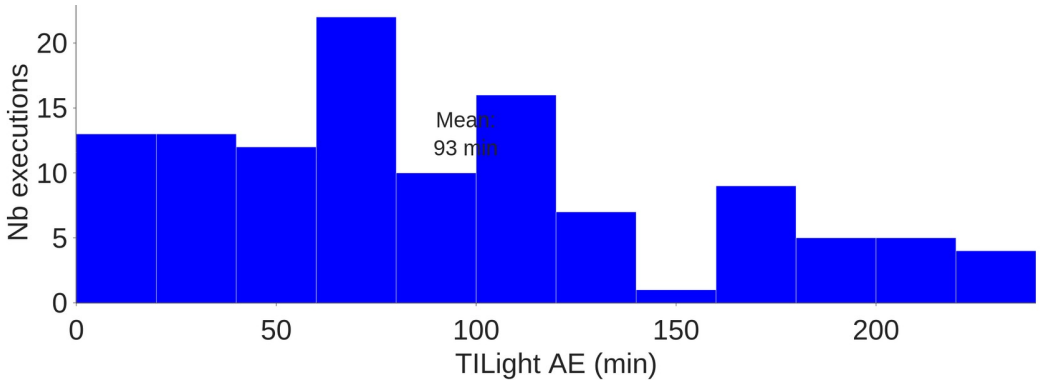


Figure S29: Histogram of TILight Absolute Error (min) on 117 subjects

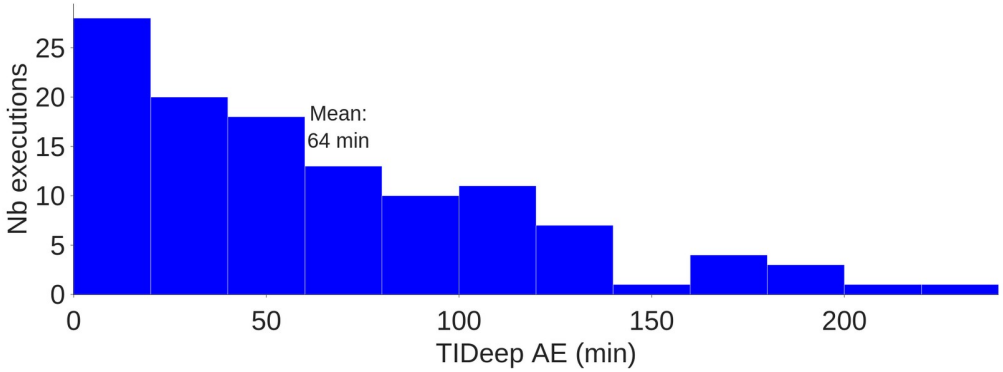


Figure S30: Histogram of TIDeep Absolute Error (min) on 117 subjects

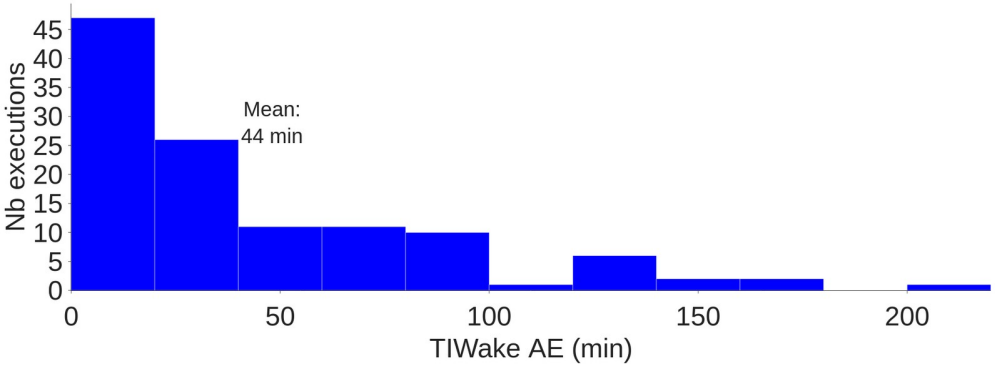


Figure S31: Histogram of TIWake Absolute Error (min) on 117 subjects

Metadata	Subgroup	Number of subjects
bmi_group	<18.5	2
mattress	unknown	2
hypnotics	yes	1
psqi_score_group	no_sleep_difficulty	1
psqi_score_group	severy_sleep_difficulty	1
sleep_quality	very_good	1

Table S5: Subgroup Distribution of Participants with Insufficient Sample Sizes for Statistical Analysis

REFERENCES

1. Berry RB, Brooks R, Gamaldo CE, Harding SM, Marcus CL, Vaughn BV. The AASM manual for the scoring of sleep and associated events. Rules, Terminology and Technical Specifications. Darien, Illinois: American Academy of Sleep Medicine; 2012. https://aasm.org/resources/pdf/scoring-manual-preface.pdf [accessed July 20, 2024]

2. Lee YJ, Lee JY, Cho JH, Choi JH. Interrater reliability of sleep stage scoring: a meta-analysis. J Clin Sleep Med American Academy of Sleep Medicine; 18(1):193–202. doi: 10.5664/jcsm.9538

3. Miles J. R-squared, adjusted R-squared. Encycl Stat Behav Sci 2005. doi:10.1002/0470013192.bsa526

4. Dahiru T. P-value, a true test of statistical significance? A cautionary note. Ann Ib Postgrad Med 2008;6(1):21–26. PMID: 25161440

5. Derpanis KG. Overview of the RANSAC Algorithm. Image Rochester NY 2010;4(1):2–3. https://api.semanticscholar.org/CorpusID:2948081 [accessed Sept 18, 2024]

B.2 Results of Kruskal-Wallis and Brown-Forsythe statistical tests followed by Holm-Bonferroni correction

hypothesis was rejected, indicating a significant difference in variances.

metrics	metadata	alpha	k_stat	k_p_value	b_stat	b_p_value	adjusted_k_alpha	k_rejected	b_p_values	adjusted_b_alpha	b_rejected
TST	bmi_group	0.05	3.7E+00	3.0E-01	5.3E-01	6.7E-01	3.3E-04	0.00E+00	6.7E-01	6.8E-04	0.00E+00
TST	sex	0.05	3.8E-01	5.4E-01	1.0E+00	3.1E-01	5.8E-04	0.00E+00	3.1E-01	2.9E-04	0.00E+00
TST	psqi_score_group	0.05	8.0E+00	4.6E-02	2.2E+00	9.0E-02	2.3E-04	0.00E+00	9.0E-02	2.3E-04	0.00E+00
TST	psqi_quality_group	0.05	4.5E+00	3.4E-02	3.7E+00	5.7E-02	2.2E-04	0.00E+00	5.7E-02	2.2E-04	0.00E+00
TST	psqi_latency_group	0.05	3.9E+00	4.9E-02	6.6E+00	1.1E-02	2.3E-04	0.00E+00	1.1E-02	2.1E-04	0.00E+00
TST	psqi_duration_group	0.05	1.4E+01	1.9E-04	2.0E+00	1.6E-01	2.1E-04	**1.00E+00**	1.6E-01	2.5E-04	0.00E+00
TST	psqi_efficiency_group	0.05	2.7E+00	9.8E-02	4.2E-01	5.2E-01	2.6E-04	0.00E+00	5.2E-01	4.6E-04	0.00E+00
TST	psqi_disturbances_group	0.05	7.4E-01	3.9E-01	8.9E-02	7.7E-01	4.2E-04	0.00E+00	7.7E-01	1.0E-03	0.00E+00
TST	psqi_medication_group	0.05	1.6E-01	6.9E-01	2.3E-01	6.3E-01	9.6E-04	0.00E+00	6.3E-01	6.3E-04	0.00E+00
TST	psqi_day_dysfunction_group	0.05	5.2E+00	2.3E-02	1.8E-01	6.7E-01	2.2E-04	0.00E+00	6.7E-01	7.0E-04	0.00E+00
TST	mattress	0.05	1.5E+00	9.1E-01	1.1E+00	3.9E-01	3.6E-03	0.00E+00	3.9E-01	3.4E-04	0.00E+00
TST	mattress_thickness_group	0.05	3.1E+00	7.6E-02	6.9E-03	9.3E-01	2.4E-04	0.00E+00	9.3E-01	3.6E-03	0.00E+00
TST	is_couple	0.05	1.1E-03	9.7E-01	2.1E-03	9.6E-01	8.3E-03	0.00E+00	9.6E-01	7.1E-03	0.00E+00
TST	age_group	0.05	5.6E+00	6.0E-02	8.7E-01	4.2E-01	2.4E-04	0.00E+00	4.2E-01	3.6E-04	0.00E+00
TILight	bmi_group	0.05	9.2E+00	2.7E-02	7.2E-01	5.4E-01	2.2E-04	0.00E+00	5.4E-01	4.7E-04	0.00E+00
TILight	sex	0.05	3.1E+00	7.8E-02	3.4E-02	8.5E-01	2.4E-04	0.00E+00	8.5E-01	1.8E-03	0.00E+00
TILight	psqi_score_group	0.05	2.7E+00	4.3E-01	1.1E+00	3.6E-01	4.7E-04	0.00E+00	3.6E-01	3.3E-04	0.00E+00
TILight	psqi_quality_group	0.05	1.4E+00	2.4E-01	2.2E-04	9.9E-01	3.0E-04	0.00E+00	9.9E-01	1.7E-02	0.00E+00
TILight	psqi_latency_group	0.05	5.6E-02	8.1E-01	1.5E-02	9.0E-01	1.7E-03	0.00E+00	9.0E-01	2.6E-03	0.00E+00
TILight	psqi_duration_group	0.05	2.2E-01	6.4E-01	4.7E-01	5.0E-01	8.1E-04	0.00E+00	5.0E-01	4.2E-04	0.00E+00
TILight	psqi_efficiency_group	0.05	2.0E+00	1.6E-01	8.4E-01	3.6E-01	2.7E-04	0.00E+00	3.6E-01	3.2E-04	0.00E+00
TILight	psqi_disturbances_group	0.05	3.5E+00	6.0E-02	5.4E-01	4.6E-01	2.4E-04	0.00E+00	4.6E-01	4.1E-04	0.00E+00
TILight	psqi_medication_group	0.05	1.1E+00	3.0E-01	6.7E-01	4.1E-01	3.4E-04	0.00E+00	4.1E-01	3.6E-04	0.00E+00
TILight	psqi_day_dysfunction_group	0.05	4.0E-02	8.4E-01	2.0E-01	6.6E-01	1.9E-03	0.00E+00	6.6E-01	6.7E-04	0.00E+00
TILight	mattress	0.05	2.0E+01	1.4E-03	1.1E+00	3.8E-01	2.1E-04	0.00E+00	3.8E-01	3.4E-04	0.00E+00
TILight	mattress_thickness_group	0.05	1.3E-01	7.2E-01	4.8E+00	3.0E-02	1.1E-03	0.00E+00	3.0E-02	2.1E-04	0.00E+00
TILight	is_couple	0.05	7.9E-01	3.7E-01	6.3E-01	4.3E-01	4.0E-04	0.00E+00	4.3E-01	3.7E-04	0.00E+00
TILight	age_group	0.05	2.1E+00	3.6E-01	6.6E-01	5.2E-01	3.8E-04	0.00E+00	5.2E-01	4.5E-04	0.00E+00
TIDeep	bmi_group	0.05	2.2E+00	5.3E-01	1.3E+00	2.7E-01	5.6E-04	0.00E+00	2.7E-01	2.8E-04	0.00E+00
TIDeep	sex	0.05	3.6E-01	5.5E-01	3.7E+00	5.6E-02	6.0E-04	0.00E+00	5.6E-02	2.2E-04	0.00E+00
TIDeep	psqi_score_group	0.05	2.9E+00	4.1E-01	1.2E+00	3.0E-01	4.5E-04	0.00E+00	3.0E-01	2.9E-04	0.00E+00
TIDeep	psqi_quality_group	0.05	1.7E-01	6.8E-01	9.7E-02	7.6E-01	9.1E-04	0.00E+00	7.6E-01	9.6E-04	0.00E+00
TIDeep	psqi_latency_group	0.05	7.2E-02	7.9E-01	2.4E-04	9.9E-01	1.3E-03	0.00E+00	9.9E-01	1.3E-02	0.00E+00
TIDeep	psqi_duration_group	0.05	5.8E-02	8.1E-01	1.6E-01	6.9E-01	1.6E-03	0.00E+00	6.9E-01	7.5E-04	0.00E+00
TIDeep	psqi_efficiency_group	0.05	2.4E-01	6.2E-01	2.2E-01	6.4E-01	7.1E-04	0.00E+00	6.4E-01	6.3E-04	0.00E+00
TIDeep	psqi_disturbances_group	0.05	2.4E+00	1.2E-01	6.0E-01	4.4E-01	2.7E-04	0.00E+00	4.4E-01	3.8E-04	0.00E+00
TIDeep	psqi_medication_group	0.05	2.0E+00	1.6E-01	8.4E-01	3.6E-01	2.8E-04	0.00E+00	3.6E-01	3.3E-04	0.00E+00
TIDeep	psqi_day_dysfunction_group	0.05	2.7E-03	9.6E-01	3.1E-01	5.8E-01	6.3E-03	0.00E+00	5.8E-01	5.2E-04	0.00E+00
TIDeep	mattress	0.05	1.4E+01	1.8E-02	2.4E+00	4.4E-02	2.2E-04	0.00E+00	4.4E-02	2.2E-04	0.00E+00
TIDeep	mattress_thickness_group	0.05	1.5E+00	2.2E-01	6.8E-02	8.0E-01	3.0E-04	0.00E+00	8.0E-01	1.1E-03	0.00E+00
TIDeep	is_couple	0.05	1.6E+00	2.0E-01	1.4E+00	2.5E-01	2.9E-04	0.00E+00	2.5E-01	2.6E-04	0.00E+00
TIDeep	age_group	0.05	7.2E-01	7.0E-01	1.1E+00	3.4E-01	9.8E-04	0.00E+00	3.4E-01	3.2E-04	0.00E+00
TIREM	bmi_group	0.05	5.2E+00	1.6E-01	8.9E-01	4.5E-01	2.7E-04	0.00E+00	4.5E-01	3.9E-04	0.00E+00
TIREM	sex	0.05	7.7E-04	9.8E-01	2.0E-01	6.6E-01	1.0E-02	0.00E+00	6.6E-01	6.8E-04	0.00E+00
TIREM	psqi_score_group	0.05	1.7E+00	6.3E-01	1.1E+00	3.7E-01	7.4E-04	0.00E+00	3.7E-01	3.4E-04	0.00E+00
TIREM	psqi_quality_group	0.05	3.6E-01	5.5E-01	3.1E-04	9.9E-01	6.1E-04	0.00E+00	9.9E-01	1.0E-02	0.00E+00
TIREM	psqi_latency_group	0.05	1.3E+00	2.5E-01	2.5E-01	6.2E-01	3.1E-04	0.00E+00	6.2E-01	6.0E-04	0.00E+00
TIREM	psqi_duration_group	0.05	2.9E+00	8.7E-02	7.1E-05	9.9E-01	2.5E-04	0.00E+00	9.9E-01	2.5E-02	0.00E+00
TIREM	psqi_efficiency_group	0.05	2.9E+00	8.8E-02	2.3E+00	1.3E-01	2.5E-04	0.00E+00	1.3E-01	2.4E-04	0.00E+00
TIREM	psqi_disturbances_group	0.05	1.7E-02	9.0E-01	2.5E-01	6.2E-01	3.1E-03	0.00E+00	6.2E-01	6.0E-04	0.00E+00
TIREM	psqi_medication_group	0.05	4.6E-01	5.0E-01	2.8E-01	6.0E-01	5.3E-04	0.00E+00	6.0E-01	5.6E-04	0.00E+00
TIREM	psqi_day_dysfunction_group	0.05	2.2E-01	6.4E-01	3.0E-01	5.9E-01	7.9E-04	0.00E+00	5.9E-01	5.4E-04	0.00E+00
TIREM	mattress	0.05	1.1E+00	9.6E-01	3.4E-01	8.9E-01	5.6E-03	0.00E+00	8.9E-01	2.3E-03	0.00E+00
TIREM	mattress_thickness_group	0.05	4.5E+00	3.4E-02	1.5E+00	2.2E-01	2.2E-04	0.00E+00	2.2E-01	2.6E-04	0.00E+00
TIREM	is_couple	0.05	1.2E-04	9.9E-01	1.5E+00	2.2E-01	1.7E-02	0.00E+00	2.2E-01	2.6E-04	0.00E+00
TIREM	age_group	0.05	2.8E+00	2.5E-01	5.7E-01	5.7E-01	3.1E-04	0.00E+00	5.7E-01	5.1E-04	0.00E+00
TIWake	bmi_group	0.05	3.7E+00	3.0E-01	5.3E-01	6.7E-01	3.38E-04	0.0E+00	6.7E-01	7.14E-04	0.00E+00
TIWake	sex	0.05	3.8E-01	5.4E-01	1.0E+00	3.1E-01	5.32E-04	0.0E+00	3.1E-01	2.78E-04	0.00E+00
TIWake	psqi_score_group	0.05	8.0E+00	4.6E-02	2.2E+00	9.0E-02	2.30E-04	0.0E+00	9.0E-02	2.22E-04	0.00E+00
TIWake	psqi_quality_group	0.05	4.5E+00	3.4E-02	3.7E+00	5.7E-02	2.24E-04	0.0E+00	5.7E-02	2.16E-04	0.00E+00
TIWake	psqi_latency_group	0.05	3.9E+00	4.9E-02	6.6E+00	1.1E-02	2.33E-04	0.0E+00	1.1E-02	2.11E-04	0.00E+00
TIWake	psqi_duration_group	0.05	1.4E+01	1.9E-04	2.0E+00	1.6E-01	2.10E-04	**1.0E+00**	1.6E-01	2.35E-04	0.00E+00
TIWake	psqi_efficiency_group	0.05	2.7E+00	9.8E-02	4.2E-01	5.2E-01	2.59E-04	0.0E+00	5.2E-01	4.03E-04	0.00E+00
TIWake	psqi_disturbances_group	0.05	7.4E-01	3.9E-01	8.9E-02	7.7E-01	4.07E-04	0.0E+00	7.7E-01	1.22E-03	0.00E+00
TIWake	psqi_medication_group	0.05	1.6E-01	6.9E-01	2.3E-01	6.3E-01	7.58E-04	0.0E+00	6.3E-01	5.81E-04	0.00E+00
TIWake	psqi_day_dysfunction_group	0.05	5.2E+00	2.3E-02	1.8E-01	6.7E-01	2.19E-04	0.0E+00	6.7E-01	7.46E-04	0.00E+00
TIWake	mattress	0.05	1.5E+00	9.1E-01	1.1E+00	3.9E-01	2.50E-03	0.0E+00	3.9E-01	3.18E-04	0.00E+00
TIWake	mattress_thickness_group	0.05	3.1E+00	7.6E-02	6.9E-03	9.3E-01	2.48E-04	0.0E+00	9.3E-01	3.33E-03	0.00E+00
TIWake	is_couple	0.05	1.1E-03	9.7E-01	2.1E-03	9.6E-01	8.33E-03	0.0E+00	9.6E-01	7.14E-03	0.00E+00
TIWake	hypnotics_night1	0.05	1.0E-01	7.5E-01	2.3E-01	6.3E-01	9.26E-04	0.0E+00	6.3E-01	5.95E-04	0.00E+00
TIWake	sleep_quality_night1	0.05	5.4E+00	2.5E-01	5.8E-01	6.8E-01	3.16E-04	0.0E+00	6.8E-01	7.69E-04	0.00E+00
TIWake	age_group	0.05	5.6E+00	6.0E-02	8.7E-01	4.2E-01	2.42E-04	0.0E+00	4.2E-01	3.36E-04	0.00E+00
PILight	bmi_group	0.05	1.2E+01	9.2E-03	5.9E-01	6.2E-01	2.14E-04	0.0E+00	6.2E-01	5.56E-04	0.00E+00
PILight	sex	0.05	2.0E+00	1.5E-01	1.8E-05	1.0E+00	2.79E-04	0.0E+00	1.0E+00	5.00E-02	0.00E+00
PILight	psqi_score_group	0.05	2.5E+00	4.8E-01	1.2E+00	3.1E-01	5.00E-04	0.0E+00	3.1E-01	2.75E-04	0.00E+00
PILight	psqi_quality_group	0.05	8.6E-01	3.5E-01	3.7E-02	8.5E-01	3.82E-04	0.0E+00	8.5E-01	1.72E-03	0.00E+00
PILight	psqi_latency_group	0.05	5.6E-02	8.1E-01	2.2E-01	6.4E-01	1.19E-03	0.0E+00	6.4E-01	6.25E-04	0.00E+00
PILight	psqi_duration_group	0.05	7.0E-01	4.0E-01	1.3E+00	2.6E-01	4.20E-04	0.0E+00	2.6E-01	2.54E-04	0.00E+00

PILight	psqi_efficiency_group	0.05	1.8E+00	1.8E-01	6.2E-01	4.3E-01	2.89E-04	0.0E+00	4.3E-01	3.47E-04	0.00E+00
PILight	psqi_disturbances_group	0.05	3.3E+00	7.0E-02	9.4E-02	7.6E-01	2.45E-04	0.0E+00	7.6E-01	1.16E-03	0.00E+00
PILight	psqi_medication_group	0.05	1.2E+00	2.8E-01	5.6E-01	4.6E-01	3.29E-04	0.0E+00	4.6E-01	3.68E-04	0.00E+00
PILight	psqi_day_dysfunction_group	0.05	5.4E-03	9.4E-01	1.1E-01	7.5E-01	3.85E-03	0.0E+00	7.5E-01	1.02E-03	0.00E+00
PILight	mattress	0.05	1.9E+01	2.3E-03	1.2E+00	3.0E-01	2.13E-04	0.0E+00	3.0E-01	2.67E-04	0.00E+00
PILight	mattress_thickness_group	0.05	1.0E-01	7.5E-01	5.0E+00	2.7E-02	9.43E-04	0.0E+00	2.7E-02	2.12E-04	0.00E+00
PILight	is_couple	0.05	1.1E+00	2.9E-01	6.4E-01	4.3E-01	3.31E-04	0.0E+00	4.3E-01	3.42E-04	0.00E+00
PILight	hypnotics_night1	0.05	3.7E-03	9.5E-01	1.5E+00	2.3E-01	4.55E-03	0.0E+00	2.3E-01	2.48E-04	0.00E+00
PILight	sleep_quality_night1	0.05	2.2E+00	7.0E-01	1.4E+00	2.3E-01	7.81E-04	0.0E+00	2.3E-01	2.49E-04	0.00E+00
PILight	age_group	0.05	2.8E+00	2.4E-01	1.9E+00	1.5E-01	3.09E-04	0.0E+00	1.5E-01	2.34E-04	0.00E+00
PIDeep	bmi_group	0.05	2.8E+00	4.2E-01	1.1E+00	3.6E-01	4.27E-04	0.0E+00	3.6E-01	2.99E-04	0.00E+00
PIDeep	sex	0.05	1.7E-01	6.8E-01	3.1E+00	8.0E-02	7.35E-04	0.0E+00	8.0E-02	2.18E-04	0.00E+00
PIDeep	psqi_score_group	0.05	2.8E+00	4.2E-01	1.2E+00	3.2E-01	4.35E-04	0.0E+00	3.2E-01	2.81E-04	0.00E+00
PIDeep	psqi_quality_group	0.05	5.3E-02	8.2E-01	1.3E-01	7.2E-01	1.22E-03	0.0E+00	7.2E-01	8.62E-04	0.00E+00
PIDeep	psqi_latency_group	0.05	2.4E-02	8.8E-01	3.9E-02	8.4E-01	1.67E-03	0.0E+00	8.4E-01	1.56E-03	0.00E+00
PIDeep	psqi_duration_group	0.05	2.2E-02	8.8E-01	7.2E-03	9.3E-01	1.92E-03	0.0E+00	9.3E-01	2.94E-03	0.00E+00
PIDeep	psqi_efficiency_group	0.05	1.3E-01	7.1E-01	3.2E-01	5.7E-01	8.06E-04	0.0E+00	5.7E-01	4.63E-04	0.00E+00
PIDeep	psqi_disturbances_group	0.05	2.6E+00	1.0E-01	2.8E-01	6.0E-01	2.60E-04	0.0E+00	6.0E-01	5.10E-04	0.00E+00
PIDeep	psqi_medication_group	0.05	1.6E+00	2.1E-01	4.8E-01	4.9E-01	2.98E-04	0.0E+00	4.9E-01	3.82E-04	0.00E+00
PIDeep	psqi_day_dysfunction_group	0.05	2.9E-02	8.6E-01	2.0E-01	6.5E-01	1.52E-03	0.0E+00	6.5E-01	6.67E-04	0.00E+00
PIDeep	mattress	0.05	1.4E+01	1.9E-02	2.5E+00	3.4E-02	2.16E-04	0.0E+00	3.4E-02	2.14E-04	0.00E+00
PIDeep	mattress_thickness_group	0.05	1.1E+00	3.0E-01	8.9E-01	3.5E-01	3.50E-04	0.0E+00	3.5E-01	2.91E-04	0.00E+00
PIDeep	is_couple	0.05	2.5E+00	1.2E-01	1.0E+00	3.2E-01	2.72E-04	0.0E+00	3.2E-01	2.79E-04	0.00E+00
PIDeep	hypnotics_night1	0.05	1.3E-02	9.1E-01	2.6E-01	6.1E-01	2.78E-03	0.0E+00	6.1E-01	5.38E-04	0.00E+00
PIDeep	sleep_quality_night1	0.05	2.4E+00	6.7E-01	5.8E-01	6.8E-01	7.14E-04	0.0E+00	6.8E-01	7.81E-04	0.00E+00
PIDeep	age_group	0.05	8.2E-01	6.6E-01	8.3E-01	4.4E-01	7.04E-04	0.0E+00	4.4E-01	3.50E-04	0.00E+00
PIREM	bmi_group	0.05	4.1E+00	2.5E-01	5.6E-01	6.4E-01	3.11E-04	0.0E+00	6.4E-01	6.17E-04	0.00E+00
PIREM	sex	0.05	5.9E-02	8.1E-01	4.8E-03	9.4E-01	1.11E-03	0.0E+00	9.4E-01	4.55E-03	0.00E+00
PIREM	psqi_score_group	0.05	1.9E+00	5.8E-01	9.5E-01	4.2E-01	5.75E-04	0.0E+00	4.2E-01	3.29E-04	0.00E+00
PIREM	psqi_quality_group	0.05	2.2E-01	6.4E-01	7.3E-02	7.9E-01	6.41E-04	0.0E+00	7.9E-01	1.32E-03	0.00E+00
PIREM	psqi_latency_group	0.05	1.4E+00	2.4E-01	2.9E-01	5.9E-01	3.07E-04	0.0E+00	5.9E-01	5.00E-04	0.00E+00
PIREM	psqi_duration_group	0.05	3.8E+00	5.1E-02	5.7E-01	4.5E-01	2.35E-04	0.0E+00	4.5E-01	3.60E-04	0.00E+00
PIREM	psqi_efficiency_group	0.05	2.6E+00	1.1E-01	8.8E-01	3.5E-01	2.67E-04	0.0E+00	3.5E-01	2.92E-04	0.00E+00
PIREM	psqi_disturbances_group	0.05	1.4E-02	9.1E-01	1.0E-01	7.5E-01	2.63E-03	0.0E+00	7.5E-01	1.04E-03	0.00E+00
PIREM	psqi_medication_group	0.05	5.1E-01	4.8E-01	2.9E-01	5.9E-01	4.90E-04	0.0E+00	5.9E-01	4.95E-04	0.00E+00
PIREM	psqi_day_dysfunction_group	0.05	9.7E-02	7.6E-01	2.8E-01	6.0E-01	9.62E-04	0.0E+00	6.0E-01	5.05E-04	0.00E+00
PIREM	mattress	0.05	9.7E-01	9.6E-01	4.0E-01	8.5E-01	6.25E-03	0.0E+00	8.5E-01	1.79E-03	0.00E+00
PIREM	mattress_thickness_group	0.05	4.1E+00	4.3E-02	1.5E+00	2.2E-01	2.27E-04	0.0E+00	2.2E-01	2.44E-04	0.00E+00
PIREM	is_couple	0.05	4.5E-02	8.3E-01	1.4E+00	2.4E-01	1.25E-03	0.0E+00	2.4E-01	2.51E-04	0.00E+00
PIREM	hypnotics_night1	0.05	3.2E-01	5.7E-01	4.3E-02	8.4E-01	5.68E-04	0.0E+00	8.4E-01	1.52E-03	0.00E+00
PIREM	sleep_quality_night1	0.05	3.5E+00	4.7E-01	9.9E-01	4.2E-01	4.81E-04	0.0E+00	4.2E-01	3.31E-04	0.00E+00
PIREM	age_group	0.05	1.5E+00	4.8E-01	9.6E-01	3.9E-01	4.95E-04	0.0E+00	3.9E-01	3.14E-04	0.00E+00
PIWake	bmi_group	0.05	4.3E+00	2.3E-01	6.7E-01	5.7E-01	3.01E-04	0.0E+00	5.7E-01	4.67E-04	0.00E+00
PIWake	sex	0.05	1.1E-01	7.4E-01	7.6E-01	3.8E-01	8.77E-04	0.0E+00	3.8E-01	3.13E-04	0.00E+00
PIWake	psqi_score_group	0.05	8.0E+00	4.5E-02	2.2E+00	9.4E-02	2.28E-04	0.0E+00	9.4E-02	2.23E-04	0.00E+00
PIWake	psqi_quality_group	0.05	4.2E+00	4.0E-02	2.7E+00	1.0E-01	2.25E-04	0.0E+00	1.0E-01	2.25E-04	0.00E+00
PIWake	psqi_latency_group	0.05	3.7E+00	5.5E-02	7.5E+00	7.3E-03	2.38E-04	0.0E+00	7.3E-03	2.09E-04	0.00E+00
PIWake	psqi_duration_group	0.05	1.5E+01	9.3E-05	2.6E+00	1.1E-01	2.08E-04	**1.0E+00**	1.1E-01	2.26E-04	0.00E+00
PIWake	psqi_efficiency_group	0.05	2.6E+00	1.1E-01	5.2E-01	4.7E-01	2.69E-04	0.0E+00	4.7E-01	3.79E-04	0.00E+00
PIWake	psqi_disturbances_group	0.05	6.2E-01	4.3E-01	1.1E-01	7.4E-01	4.39E-04	0.0E+00	7.4E-01	1.00E-03	0.00E+00
PIWake	psqi_medication_group	0.05	1.6E-01	6.9E-01	1.0E-01	7.5E-01	7.69E-04	0.0E+00	7.5E-01	1.06E-03	0.00E+00
PIWake	psqi_day_dysfunction_group	0.05	5.2E+00	2.2E-02	2.3E-01	6.3E-01	2.17E-04	0.0E+00	6.3E-01	5.68E-04	0.00E+00
PIWake	mattress	0.05	1.7E+00	8.9E-01	1.1E+00	3.7E-01	2.08E-03	0.0E+00	3.7E-01	3.09E-04	0.00E+00
PIWake	mattress_thickness_group	0.05	3.4E+00	6.5E-02	4.8E-03	9.4E-01	2.43E-04	0.0E+00	9.4E-01	5.00E-03	0.00E+00
PIWake	is_couple	0.05	6.0E-02	8.1E-01	1.6E-01	6.9E-01	1.09E-03	0.0E+00	6.9E-01	7.94E-04	0.00E+00
PIWake	hypnotics_night1	0.05	2.1E-01	6.4E-01	2.0E-01	6.6E-01	6.85E-04	0.0E+00	6.6E-01	6.76E-04	0.00E+00
PIWake	sleep_quality_night1	0.05	5.1E+00	2.8E-01	7.6E-01	5.5E-01	3.25E-04	0.0E+00	5.5E-01	4.24E-04	0.00E+00
PIWake	age_group	0.05	4.5E+00	1.1E-01	7.6E-01	4.7E-01	2.65E-04	0.0E+00	4.7E-01	3.76E-04	0.00E+00

B.3 Comparison with previous research

LegendEBE: Epoch by epochS-W: Sleep-Wake

Results must be interpreted with cautious as the dataset used for WSA performance evaluation has been collected in real-life conditions, is larger and more diverse than the other datasets according to the protocol described in other studies.

For all these reasons, the reader must keep in mind that the evaluation of the other devices on the associated datasets may overestimate their performance compared to the results obtained for WSA we provide in this study.

The performance presented in this document of the listed devices has been computed based on the confusion matrix when it was available, since the data for each night were not provided. This explains the discrepancies between the performance of Withings Sleep Analyzer in this document and in the full-text.

General Information							EBE metrics: S-W						EBE metrics: sleep stages				
Reference	**Devices**	**Wearable/Nearable**	**# subjects**	**Reference**	**Study setting**	**# nights**	**Comment**	**Mean reference TST (in minutes) ± std**	**Se sleep / Sp wake**	**Sp sleep / Se wake**	**Accuracy**	**Proportion of wake data (%)**	**Kappa**	**Se light**	**Se deep**	**Se REM**	**Accuracy**	**Kappa**
Chinoy2021	Philips Respironics Actiwatch 2	W	34 healthy young adults	PSG (Siesta; Compumedics)	Sleep Lab3 nights per subjec	98	**Exclusions post collection** We selected a duration of <470 min (ie, <7h50 min) as the insufficient total recording time threshold whereby individual nights for a device were removed from the final summary analyses. were counterbalanced across participants.	418 ± 41	0.91	0.63	0.89	8	0.42	N/A	N/A	N/A	N/A	N/A
	Fatigue Science Readiband	W	15 healthy young adults			41		416 ± 48	0.92	0.55	0.88	13	0.43					
	Fitbit Alta HR	W	20 healthy young adults			49		425 ± 33	0.95	0.54	0.90	11	0.51	0.76	0.53	0.69	0.68	0.51
	Garmin Fenix 5S	W	11 healthy young adults			29		413 ± 53	0.99	0.18	0.88	14	0.25	0.68	0.56	0.50	0.55	0.30
	Garmin Vivosmart 3	W	15 healthy young adults			43		415 ± 48	0.99	0.19	0.88	14	0.26	0.70	0.56	0.54	0.58	0.33
	EarlySense Live	N	19 healthy young adults			51		422 ± 35	0.96	0.47	0.90	12	0.48	0.57	0.68	0.64	0.59	0.40
	ResMed S+	N	19 healthy young adults			51		422 ± 34	0.93	0.51	0.88	12	0.44	0.67	0.59	0.50	0.60	0.39
	SleepScore Max	N	15 healthy young adults			42		414 ± 49	0.94	0.50	0.88	14	0.46	0.68	0.59	0.49	0.60	0.39
Kainec2024	Philips Repironics Actiwatch Spectrum Plus	W	53 healthy young adults	PSG (EasyCap for 33 participants & EasyCap + BrainAmp MR plus amplifiers, Brain Products GmbH for 20 other participants)	Sleep Lab	53	**Short nights exclusion** Three participants slept less than 4.5 h overnight in the lab and were therefore excluded from further analyses. **Heteroscedasticity** Briefly, plots were examined for heteroscedasticity and proportional bias. Where heteroscedasticity and/or proportional bias	435 ± 57	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Fitbit Inspire HR	W				53		434 ± 53										
	Fitbit Versa 2	W				53		435 ± 57										
	Garmin Vivosmart 4	W				53		435 ± 54										
	Oura Ring	W				53		435 ± 54										
	Withings Sleep Analyzer	N				53		436 ± 56										
Miller2022	Apple Watch S6	W	53 healthy young adults	PSG Grass gold-cup electrodes (AstroMed)	Sleep Lab	1/participant	[illegible] sleep detection algorithms. For all such wearable devices, the self-determined sleep period was within the imposed time in bed for all participants. Any time in bed before each device's self-determined sleep onset or after each device's self-determined sleep offset was scored by the researchers as wake. In contrast, Somfit did not have an	N/A	0.96	0.26	0.73	33	0.27	0.44	0.71		0.47	0.20
	Garmin Forerunner 245	W							0.98	0.27	0.81	25	0.33	0.68	0.28	0.50	0.43	0.24
	Polar Vantage V	W							0.93	0.51	0.82	25	0.49	0.60	0.33	0.49	0.48	0.31
	Oura Ring Generation 2	W							0.94	0.57	0.85	25	0.56	0.66	0.62	0.52	0.59	0.46
	WHOOP 3.0	W							0.92	0.56	0.83	25	0.52	0.58	0.62	0.66	0.60	0.57
	Somfit	W							0.89	0.57	0.82	20	0.45	0.68	0.68	0.60	0.64	0.50
Tal2017i	EarlySense_i	N	43	Sleep System)	setup I: Sleep Lab	1/participant	volunteers with no major sleep disorder, which was confirmed by PSG. Subjects were studied for 1 night in the sleep laboratory (setup I), with one person in bed at home (setup II), or with	N/A	0.90	0.83	0.89	19	0.66	0.63	0.54	0.40	0.63	0.42
Tal2017ii	EarlySense_ii		7	The home setups included one of two models of portable/partial PSG: Embletta Gold or X100	II: at home with no sleep pa				0.95	0.72	0.92	14	0.68	0.64	0.68	0.63	0.66	0.49
Tal2017iii	EarlySense_iii		13		III: at home with a sleep pa				0.95	0.79	0.93	15	0.72	0.68	0.52	0.64	0.66	0.50
Ravindran2023	Emfit QS	N	16 community dwelling older adults	SomnoHD PSG system (SOMNOmedics GmbH)	Sleep Lab	1/participant	conditions, and consuming <28 units of alcohol per week * **Limited time-in-bed period**: 10 hours	422 ± 53	0.99	0.21	0.67	0.41	0.23	0.65	0.41	0.28	0.41	0.18
	Withings Sleep Analyzer	N	35 community dwelling older adults						0.94	0.36	0.75	0.34	0.35	0.52	0.76	0.42	0.49	0.28
	Somnofy	N	17 community dwelling older adults						0.97	0.57	0.81	0.42	0.58	0.68	0.58	0.62	0.62	0.45
Ding2022	Sleeptracker-AI Monitor	N	102 adults	Philips Respironics ALICE 6 LDXs System	Sleep Lab	1/participant	**Population**: Big sample size, large age and BMI ranges **Limits**: Lights-off period	418 ± 85	N/A	0.71	N/A	N/A	N/A	0.85	0.66	0.80	0.79	0.68
Ellender2022	Beddit Resmed S+ Jawbone UP3	N	45 adults	N/A	Sleep Lab	1/participant	Epoch by epoch data are not available	371 ± 69	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Hsiou2023	Beddit 3.0 Beddit 3.5	N	35 healthy adults	Grass Comet XL Plus system (Astro-Nova, Inc., West Warwick, RI)	Sleep Lab	1/participant	Epoch by epoch data are not available	6 for Beddit 3.5399 ± 85 for B	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Tuominen2019	Beddit	N	10 healthy young adults	Medcare Flaga hf. Medical Devices, Reykjavik, Iceland	Sleep Lab	1/participant	**No REM stage**: while available in the earlier version of the app, REM sleep was no longer available during the study PSG REM sleep did not predictably correlate with any BST stage. PSG REM epochs were dropped for the analysis. Epoch by epoch agreement is assessed but the provided confusion matrix is normalized so the metrics of this document cannot be computed from it.	413 ± 58	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.10
Kholghi2022	Emfit QS	N	33 adult patients	N/A	Sleep Lab	1/participant	**Population**: Large age and BMI ranges	303 ± 95	N/A	0.10	0.87	N/A	0.13	0.63	0.39	0.46	0.65	N/A
Manners2025	Withings Sleep Analyzer	N	22 healthy adults	Compumedics Grael 4K PSG:EEG devices (Compumedics, Victoria, Australia).	Sleep Lab	1 and 3 nights/participant	comparison are available for a subset only. The article also splits the results into nighttime and daytime datasets. To perform meaningful	436 ± 81	[0.91-0.97]	0.43	N/A	N/A	N/A	0.63	0.75	0.70	N/A	N/A
	Fitbit Charge 4	W							N/A	0.53	N/A	N/A	N/A	0.76	0.48	0.67	N/A	N/A
Our work	Withings Sleep Analyzer	N	117 healthy adults	CID-LXe (PSG without nasal cannula and oximeter)	at home	1/subject	comparisons with other existing works, we N/A	435 ± 71	0.93	0.73	0.87	29	0.63	0.68	0.53	0.54	0.63	0.50

Appendix C

Supplementary material for Chapter 4 (SLEEPI data-collection protocol)

C.1 Identification and assessment of data loss sources

Identification and assessment of data loss sources

This document details the calculations leading to the expected number of usable inclusions, meaning the inclusions for which more than 60% of the expected objective data (devices) and60% of the expected subjective data (questionnaires) is collect

SUMMARY:Based on this prospective attrition model, an initial outreach to approximately 85,000 individuals is required to yield a final cohort of 1,000 participants with usable data, a sample size that aligns with the requirements determined by

The following table provides identified sources of data loss relevant for this study.
For each risk of data loss, a definition, the rationale for the retention rate estimation and the final retention rate are provided.For the rates based on prior Withings internal research, the data comes from studies conducted on cohorts of 10,000 to over 1,000,000 participants.

Sources of data loss		Phase Description	Retention Rate Definition	Rationale for estimation	Retention rates
sending		The initial phase involves sending recruitment communications (e.g., in-app notifications, push notifications, emails) to a pool of active Withings users in France.	Ratio of users who will be sent a recruitment communication to the active users in Withings user database.	The number contacts is determined by this prospective attrition model, so sending rate is 100%.	1
reception		This stage accounts for the proportion of subjects who actually receive the recruitment communication, ensuring the contact method (e.g., email address) is valid, and the communication is not blocked (e.g., by spam filters).	Ratio of subjects who will actually receive the recruitment communication to the subjects to whom a recruitment communication is sent.	Based on previous results from a prior Withings internal study and on the fact that only active users will be contacted via several channels (push notifications, in-app notifications and emails). Reception, answering and acceptance rates are the same for all users, whatever the devices they own.	0.64
answering		This refers to the proportion of subjects who, having received the communication, interact with it by opening it and clicking on the study link (e.g., to an initial questionnaire or information page).	Ratio of subjects who answers the recruitment communication to the subjects who received it		0.10
acceptance		This represents the proportion of subjects who, after reviewing initial study information, provide consent to participate in the screening phase.	Ratio of subjects who agree to take part of the study to the subjects who answered the recruitment communication.		0.77
inclusion criteria	**SW**	Participants are screened against specific inclusion criteria.	Ratio of subjects who fill inclusion criteria. It has been calculated using the ratio of subjects who does not enter in any exclusion criterion.	Retention rate estimation is based on the following criteria and associated rate estimation from clients tickets: - Subjects who are 18 years old or more, not vulnerable and have a French insurance. - for WSA: - how often the participant noticed data loss issues - how often another person sleeps over the participant's WSA - how often does the participant sleep the entire night over WSA - if WSA is likely to be deteriorated (damaged microphone sensor or abnormal perpetual noise coming from WSA) - for SW and SW2 - how often the participant noticed data loss issues - how often another person uses the participant's SW or SW2 As a customer ticket does not necessary mean that a device is deteriorated and several tickets can belong to the same user, so the given inclusion rate is underevaluated.	0.77
	SW2				0.94
	WSA				0.96
adherence	**SW/SW2 day-time**	Continued adherence to device usage is assessed over the 6-month active data collection period, with a requirement of device use on more than 60% of the time. Rates are device-specific: - SW/SW2 Users: Adherence is assessed for both night-time wear and day-time wear. - WSA Users: Adherence is assessed based on night-time use.	Device-related Adherence Retention Rate: Ratio of users who used their device more than 60% of the time in the last 6 months.	The estimation of device-related adherence retention rate is based on Withings user database recent use. - Note for SW2: Due to its recent launch, adherence rates for the SW2 are proxied using historical data from the SW device. Global adherence is considered as the minimum of both retention rates. - Adherence rates in this research context, with engaged participants and compensation, are anticipated to be potentially higher than historical user data. The estimation of survey-related adherence retention rate has been arbitrarily defined. Has been arbitrarily defined. This rate is estimated conservatively, as this is the first digital longitudinal study of this scale our team has conducted, and therefore lacks experience and direct benchmarks. As completing questionnaires is the primary additional burden for this cohort of active users, we assume it is the limiting factor for overall adherence; consequently, the survey adherence rate is used as a proxy for combined device-and-survey compliance.	0.60
	SW/SW2 night-time				0.64
	WSA				0.77
	surveys	Continued adherence to answering the study questionnaires is assessed over the 6-month active data collection period, with a requirement of completing more than 60% of the study surveys.	Survey-related Adherence Retention Rate: Ratio of users who completed more than 60% of the study questionnaires.		0.35
engagement & proper use		This combined stage models the risk of data loss from both improper device usage (e.g., network disconnection, incorrect sensor placement) and low participant engagement leading to perfunctory survey responses, both of which are addressed through initial participant training upon inclusion.	Engagement and Proper Use Retention Rate: A proper use means not misusing the device which could cause data loss: disconnect it from bluetooth/wifi for too long or malposition the wsm02. It is expected that a large part of the population will properly use their device. In any case, a reminder about how to properly use the device will be sent the the participants as soon as they are included.	The estimation of this retention rate has been arbitrarily defined. It is expected to be quite high since inactive and most of subjects who misuse their device will be excluded using exclusion criteria.	0.90

C.2 Prospective Attrition Model

Prospective Attrition Model

Template							
Sources of Data Loss	engagement & proper use	adherence	inclusion criteria	acceptance	answering	reception	sending
stimation of theRetention Rate (%	engagement rate M (%)	adherence rate H (%)	inclusion rate L (%)	acceptance rate C (%)	nswering rate A (%	reception rate R (%)	sending rate S (%)
Estimation of the Number of Participants at Each Stage	rgeted nb of userswith usabl	N2 = N1 / M	N3 = N2 / H	N4 = N3 / L	N5 = N4 / C	N6 = N5 / A	N7 = N6 / R

to be contacted to reach 1,000 usable data with at least one device (SW, SW2 or WSA),the following table presents an estimation of the number of users at each stage based on the use of th

Sources of Data Loss	engagement & proper use	adherence*	inclusion criteria*	acceptance	answering	reception	sending
Estimation of theRetention Rate	0.90	0.35	0.77	0.77	0.10	0.64	1
of theNumber of Participantsat	1,000	1,111	3,175	4,123	5,354	53,544	83,662

* To estimate the number of participants providing usable data from any of their wearable or nearable device, we assumed that adherence is a participant-dependent trait rather than an independent event for each device. The combined adherence rate is therefore determined by the minimum of the two device-specific rates (e.g., 77% for ScanWatch and 96% for WSA). Given that the majority of the cohort uses the original ScanWatch, we adopted the more conservative rate of 77% for this projection and for inclusion criteria retention rate.

C.3 A priori power analysis for sample-size determination

SUPPLEMENTARY MATERIALS B

This document details the methodology for the a priori power analysis conducted to determine the necessary sample size for the study's machine learning objectives. The process consists of three core stages: defining a Data Generating Process (DGP), optimizing model hyperparameters, and running the final Monte Carlo power analysis.

1. DATA GENERATING PROCESS (DGP)

To perform the simulation, a Data Generating Process (DGP) was designed to create synthetic datasets based on a multivariate logistic regression framework[1], mimicking the anticipated complexity of the real-world study data.

The process began by generating 50 latent continuous features from a multivariate normal distribution. This step was governed by a 50x50 correlation matrix with a realistic, randomized mix of weak ($|\rho| \leq 0.1$), moderate ($0.1 < |\rho| \leq 0.4$), and strong ($0.5 \leq |\rho| < 1.0$) inter-feature correlations to simulate multicollinearity. A subset of these latent variables was then transformed to create the final feature set, composed of 40 continuous and 10 four-level categorical variables.

The binary outcome was simulated to a 30% prevalence[2,3] using a logistic function. The predictive strength of the relationship between the features and the outcome was controlled by a set of 50 predefined coefficients, which were shuffled and randomly assigned to the features. This set was composed of three tiers of effect sizes: 5 strong (drawn from a uniform distribution in the range [1.2, 2.0]), 25 moderate (range [0.5, 0.8]), and 20 slight (range [0.01, 0.1]).

Finally, to ensure a rigorous test of model generalizability to unseen subjects, each simulated sample was assigned to one of two distinct groups to facilitate a Leave-One-Group-Out cross-validation during hyperparameter tuning.

2. MODEL SPECIFICATIONS AND HYPERPARAMETER TUNING

Four machine learning models were selected for evaluation: a Support Vector Machine (SVM), a Random Forest (RF), an XGBoost model, and a Neural Network (NN).

As a preliminary step, an extensive hyperparameter tuning process was conducted for each model architecture to identify the optimal set of parameters for this specific data structure. This tuning was performed using *GridSearchCV* on a dedicated synthetic dataset of 500 samples. The grids of parameters explored and the resulting optimal set of hyperparameters for each model are detailed in Table B1.

Model	Parameter	Values Explored	Optimal Value
SVM	kernel	rbf, linear	rbf
	C	1, 10, 100	100
	gamma	0.1, 0.01, scale	0.01
Random Forest	n_estimators	100, 200	100
	max_depth	10, 20, None	10
	min_samples_split	2, 5	2
	min_samples_leaf	1, 2	2
	max_features	sqrt, 0.5	sqrt
	bootstrap	TRUE	TRUE
XGBoost	n_estimators	100, 200	200
	learning_rate	0.05, 0.1	0.1
	max_depth	3, 5, 7	3
	subsample	0.8, 1.0	0.8
	colsample_bytree	0.8, 1.0	1
	gamma	0, 0.1	0
	reg_alpha	0, 1	0
	reg_lambda	1, 10	1
Neural Network	hidden_layer_sizes	[50], [100], [50, 30]	[100]
	activation	relu	relu
	solver	adam	adam
	alpha	0.0001, 0.001	0.0001
	learning_rate_init	0.001	0.001
	early_stopping	TRUE	TRUE
	validation_fraction	0.1	0.1

Table B1: Model Architectures, Hyperparameter Tuning Grids, and Optimal Values

3. Monte Carlo Power Analysis

Using the optimal hyperparameters identified above, the main power analysis was conducted via Monte Carlo simulation. The simulation was executed for a series of eight distinct sample sizes — 100, 200, 400, 800, 1,600, 3,200, 6,400, and 12,800 — with a total of 200 simulation loops performed for each.

In each loop, a new dataset was generated via the DGP, and a model was trained using its pre-defined optimal parameters. To assess generalization, the model's performance was evaluated on a dedicated, held-out test set of 10,000 samples, which was independently generated within each simulation loop. For a given sample size, statistical power was calculated as the proportion of the 200 simulation runs where the model met a predefined success criterion, such as achieving an Area Under the Curve (AUC) of at least 0.90, 0.92, or 0.94.

The AUC is a measure of the model's discrimination. For the clinical interpretability, we also calculated the sensitivity and specificity for each model. This requires applying a classification threshold (e.g., 0.5) to the model's continuous probability score to convert it into a binary prediction (insomnia or healthy). The mean accuracy, sensitivity, and specificity values for each model are detailed in Table B2.

Model	Minimum Sample Size	Mean accuracy* (std)	Mean sensitivity* (std)	Mean specificity* (std)
SVM	800	0.88 (0.01)	0.85 (0.03)	0.90 (0.02)
NN	1600	0.88 (0.01)	0.86 (0.03)	0.89 (0.02)
RF	3200	0.83 (0.02)	0.75 (0.04)	0.90 (0.01)
XGBoost	800	0.84 (0.01)	0.80 (0.03)	0.87 (0.02)

Table B2: Simulated Minimum Sample Size and Performance Metrics for Insomnia Classification
* Computed as the mean of each metric obtained on all simulations

The simulated performance metrics for our models, with accuracy ranging from 0.83 to 0.88, sensitivity from 0.75 to 0.86, and specificity from 0.87 to 0.90, are comparable to the range of those reported in the existing literature[4–8] (accuracy: [0.73, 0.98]; sensitivity: [0.76, 0.92]; specificity: [0.82, 0.91]), which confirms the potential of this study design.

The simulation results for sample efficiency were unexpected: while we anticipated a performance ranking of Random Forest, then XGBoost, then Neural Network (NN), our findings showed that XGBoost and the NN required significantly less data than the Random Forest. This outcome can be explained by the interaction between the models' core architectures and the specifics of our experimental design.

The NN's high performance at small sample sizes is likely due to its architectural suitability for the problem. Because the synthetic data was generated using a logistic regression model, the true underlying relationship between features and the outcome is smooth. The NN, being a universal function approximator, is good at learning such smooth decision boundaries. In contrast, tree-based

models like XGBoost and Random Forest create step-wise, "boxy" boundaries, which are less efficient at capturing this specific type of relationship.

Regarding tree-based models, RF model’s low performance and XGBoost high performance with small sample sizes can be attributed to the hyperparameter tuning being performed on a small, fixed sample size of 500. This process locked XGBoost into a highly efficient "weak learner" configuration (shallow trees), which is excellent at reducing bias on small datasets. Conversely, the Random Forest was locked into a "strong learner" configuration (deeper, high-variance trees) that is prone to overfitting and requires a much larger dataset to effectively reduce variance through its bagging process.

REFERENCES

1. numpy.random.Generator.multivariate_normal — NumPy v2.3 Manual. https://numpy.org/doc/stable/reference/random/generated/numpy.random.Generator.multivariate_normal.html.

2. Morin, C. M. & Jarrin, D. C. Epidemiology of Insomnia: Prevalence, Course, Risk Factors, and Public Health Burden. *Sleep Med. Clin.* **17**, 173–191 (2022).

3. Leger, D. & others. Prevalence of insomnia in a survey of 12,778 adults in France. *J. Sleep Res.* **9**, 35–42 (2000).

4. Cheng, Y.-H., Lech, M. & Wilkinson, R. H. Simultaneous Sleep Stage and Sleep Disorder Detection from Multimodal Sensors Using Deep Learning. *Sensors* **23**, (2023).

5. Kusmakar, S. *et al.* A machine learning model for multi-night actigraphic detection of chronic insomnia: development and validation of a pre-screening tool. *R. Soc. Open Sci.* **8**, 202264 (2021).

6. Hamida, S. T.-B., Ahmed, B. & Penzel, T. A Novel Insomnia Identification Method Based on Hjorth Parameters. in *2015 IEEE INTERNATIONAL SYMPOSIUM ON SIGNAL PROCESSING AND INFORMATION TECHNOLOGY (ISSPIT)* 548–552 (2015).

7. Sharma, M., Dhiman, H. S. & Acharya, U. R. Automatic identification of insomnia using optimal antisymmetric biorthogonal wavelet filter bank with ECG signals. *Comput. Biol. Med.* **131**, 104246 (2021).

8. Shahin, M. *et al.* Deep Learning and Insomnia: Assisting Clinicians With Their Diagnosis. *IEEE J. Biomed. Health Inform.* **21**, 1546–1553 (2017).

Appendix D

Supplementary material for Chapter 5 (SLEEPI data analyses)

D.1 Device-screening exits

Of the 11,341 respondents who entered the device-screening section, 457 were screened out because they had neither a usable ScanWatch nor a usable Sleep Analyzer, leaving 10,884 who passed all screening. Table D.1 gives the full two-level detail by ScanWatch failure and Sleep Analyzer trigger, with percentages computed over the respondents who entered the section.

D.2 Subjective clinical data: supplementary figures

Figures D.1 and D.2 give the comorbidity detail: the prevalence of depression and anxiety across the ISI severity tiers, and the joint co-occurrence of insomnia, depression and anxiety.

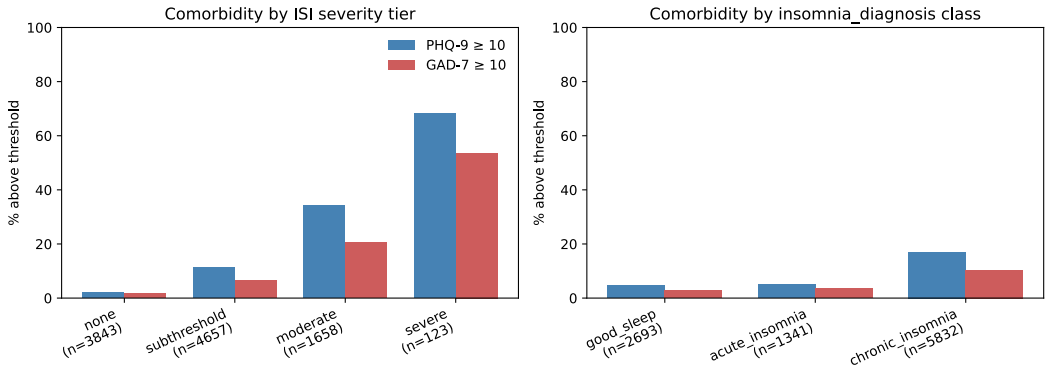


Figure D.1: Prevalence of depression (PHQ-9 $\geq$ 10) and anxiety (GAD-7 $\geq$ 10) across the ISI severity tiers.

Table D.1: Device-screening exits by ScanWatch failure and Sleep Analyzer trigger. Percentages are over the 11,341 respondents who entered the device-screening section.

ScanWatch failure	**Sleep Analyzer trigger**	n	**% opened**
No ScanWatch	No Sleep Analyzer	174	1.53
ScanWatch: frequent absent data	No Sleep Analyzer	174	1.53
ScanWatch: frequent absent data	Sleep Analyzer: frequent absent data	32	0.28
No ScanWatch	Sleep Analyzer: frequent absent data	28	0.25
No ScanWatch	Sleep Analyzer: used rarely or sometimes	14	0.12
ScanWatch: frequent absent data	Sleep Analyzer: used rarely or sometimes	10	0.09
No ScanWatch	Sleep Analyzer: often used by others	8	0.07
ScanWatch: often used by others	No Sleep Analyzer	8	0.07
No ScanWatch	Sleep Analyzer: noisy environment	2	0.02
ScanWatch: frequent absent data	Sleep Analyzer: noisy environment	2	0.02
ScanWatch: frequent absent data	Sleep Analyzer: often used by others	2	0.02
ScanWatch: often used by others	Sleep Analyzer: used rarely or sometimes	2	0.02
ScanWatch: frequent absent data	Sleep Analyzer: deteriorated microphone	1	0.01
Total exited		**457**	**4.03**

D.3 Objective feature definitions

Table D.2 defines the 39 objective features, the 31 subject-level median summaries and the 8 nonparametric circadian metrics, used for the feature description and the clustering. Each median feature is computed per night, summarized per monitoring window, and reduced to the subject median; the circadian metrics are single per-subject summaries over the whole record.

Table D.2: Definitions of the 39 objective features.

Feature	Definition	Unit
Total sleep time (TST)	Total time scored as sleep across the in-bed period.	h
Deep-sleep duration	Time scored as deep, slow-wave sleep.	h
Light-sleep duration	Time scored as light sleep.	h

continued on next page

Feature	Definition	Unit
Time in bed (TIB)	Total in-bed period from the sleep-onset attempt to the final rising.	h
Total awake duration	Duration of the awake epochs between bed-in and bed-out times.	min
Wake-up duration	Duration of the wake periods reported by the device.	min
Wake-up count	Number of wake events reported by the device per night.	count
Sleep interruptions	Number of sleep interruptions after merging short in-bed activities.	count
Sleep efficiency (SE)	Total sleep time divided by time in bed.	fraction
Sleep maintenance efficiency (SME)	Total sleep time divided by the sleep period from onset to final waking.	fraction
Awakening index	Number of wake-up events per hour of sleep.	events/h
Sleep fragmentation index	Sleep-stage transition count over the night, with wake masking applied.	index
Actigraphy fragmentation index	Fragmentation index from the ZCM activity score at threshold 5.	index
Interdaily stability (IS)	Variance of the 24-hour average activity profile over the total variance, measuring synchronization of the rhythm to the day-night cycle.	unitless
Intradaily variability (IV)	Mean square of successive hourly activity differences over the overall variance, measuring rest-activity fragmentation.	unitless
L5	Mean activity over the five least-active hours of the average day.	ZCM
M10	Mean activity over the ten most-active hours of the average day.	ZCM
Relative amplitude (RA)	$(M10 - L5)/(M10 + L5)$, the contrast between the active and rest phases.	unitless
Mean night activity	Mean ZCM activity score during the time-in-bed windows.	ZCM
Social jet lag (SJL)	Difference in mid-sleep timing between work and free days.	min
Composite phase deviation (CPD)	Deviation of daily mid-sleep from the sleep-corrected reference across work and free days.	min
Resting heart rate	Daily average resting heart rate.	bpm
Continuous heart rate	Daily average of the continuous heart-rate stream.	bpm
Minimum heart rate (sleep)	Minimum heart rate over the in-bed period.	bpm

continued on next page

Feature	Definition	Unit
Average heart rate (sleep)	Average heart rate over the in-bed period.	bpm
Maximum heart rate (sleep)	Maximum heart rate over the in-bed period.	bpm
Average SpO_2 (sleep)	Average blood-oxygen saturation over the in-bed period.	%
Active minutes	Number of minutes classed as active per day.	min/day
Inactive minutes	Number of minutes classed as inactive per day.	min/day
Light-active minutes	Number of minutes at light activity intensity per day.	min/day
Moderate-active minutes	Number of minutes at moderate activity intensity per day.	min/day
Sedentary time	Time spent inactive per day.	h
Light-intensity time	Time spent at light activity intensity per day.	min
Moderate-intensity time	Time spent at moderate activity intensity per day.	min
Vigorous-intensity time	Time spent at vigorous activity intensity per day.	min
Total activity minutes	Total minutes with an activity classification per day.	min/day
Steps	Daily step count.	steps/day
Calories burned	Daily active calories.	kcal/day
Elevation	Daily elevation gain.	m/day

D.4 Objective data: supplementary figures

Figure D.3 shows that device coverage does not decline with insomnia severity. Figures D.4, D.5 and D.6 give the feature correlation structure, the feature- and participant-wise missingness, and the distributions of the circadian metrics.

D.5 Clustering: supplementary figures

Figure D.7 reports the internal-validity scores across the clustering algorithms and cluster counts, showing that the best silhouette is low and that higher scores arise only from isolating a negligible outlier group. Figure D.8 shows the feature profile of the best available two-group partition, included to make explicit that even the most favorable solution is weak.

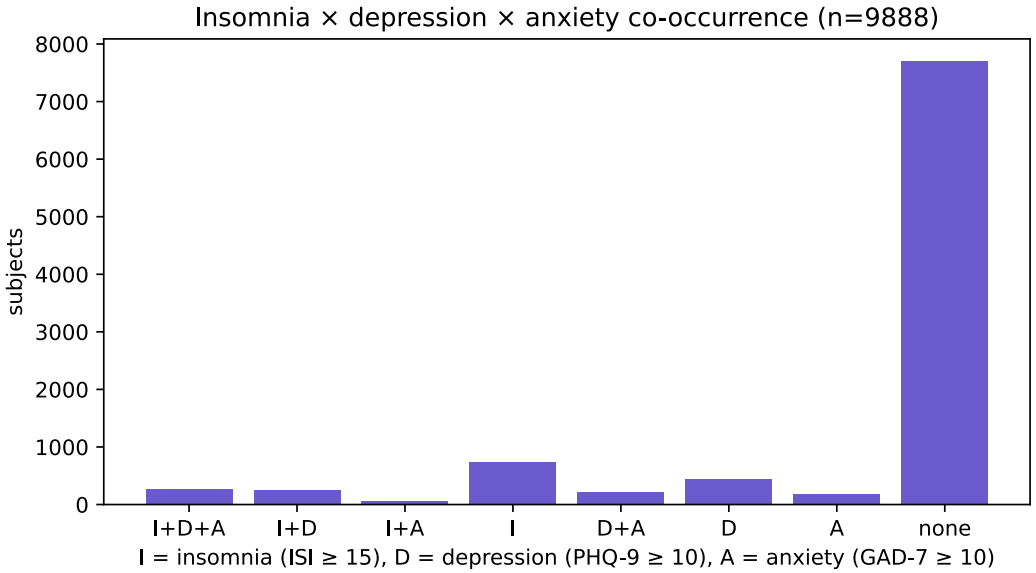


Figure D.2: Joint co-occurrence of insomnia (ISI ≥ 15), depression and anxiety among the 9,888 participants with all three scores.

D.6 Subjective–objective relationship: supplementary figures

Figure D.9 gives the full comprehensive screen of every objective feature against the three scores, the source of the strongest associations reported in the main text. Figures D.10 and D.11 report the key sleep features across the DSM-5 classes and the ISI severity tiers, where the differences are statistically significant but very small, the largest ε^2 being 0.028. Figure D.12 screens the within-person standard-deviation features against the scores, Figure D.13 gives the device-measured sleep duration and efficiency across the self-reported complaint dimension, and Figure D.14 shows the minimum-redundancy feature ranking, a descriptive ordering that picks out a small non-redundant subset of informative features and is not a validated selector.

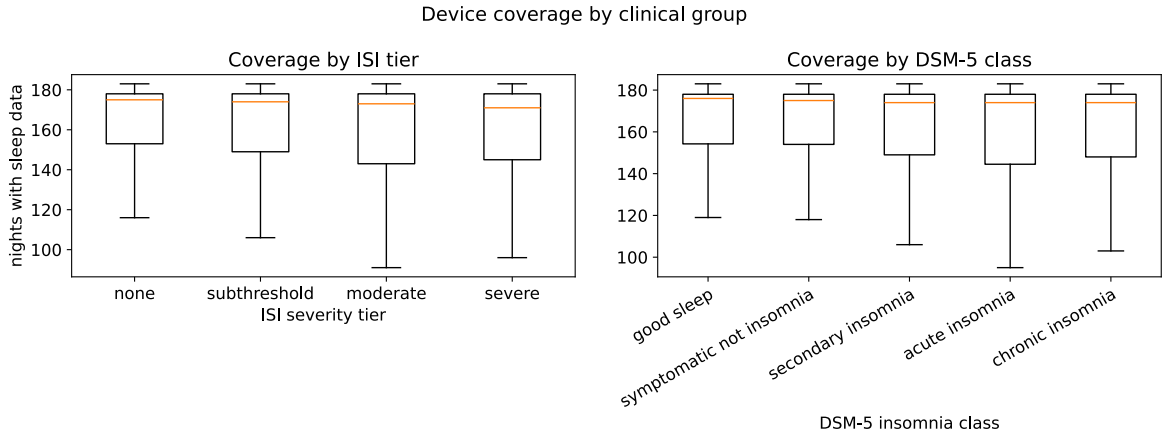


Figure D.3: Device coverage by ISI severity tier and DSM-5 insomnia class, a check that more severe insomnia is not associated with lower wear.

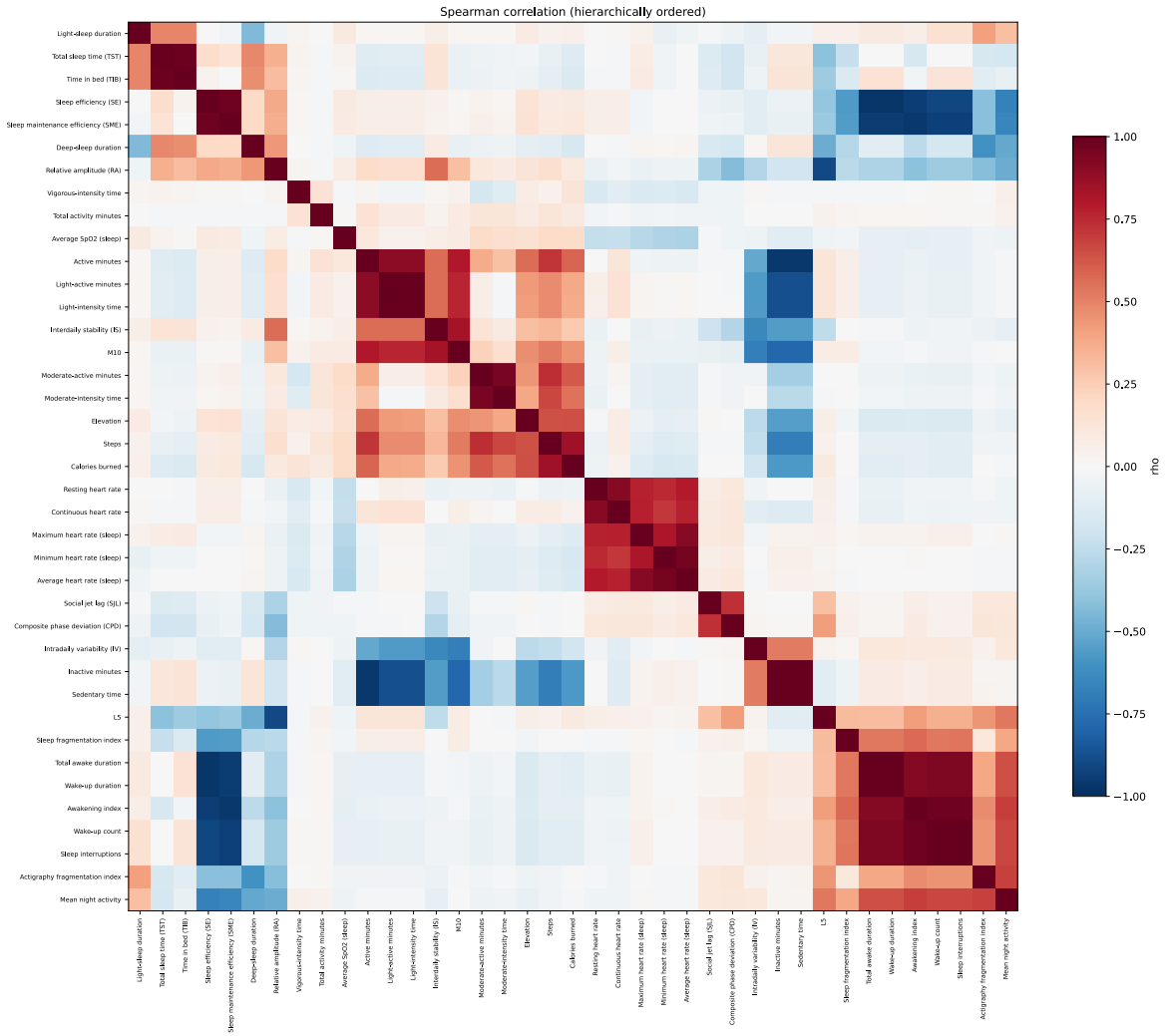


Figure D.4: Spearman correlation among the objective features, with the features reordered so that strongly correlated ones sit next to each other, making co-varying groups visible as blocks along the diagonal.

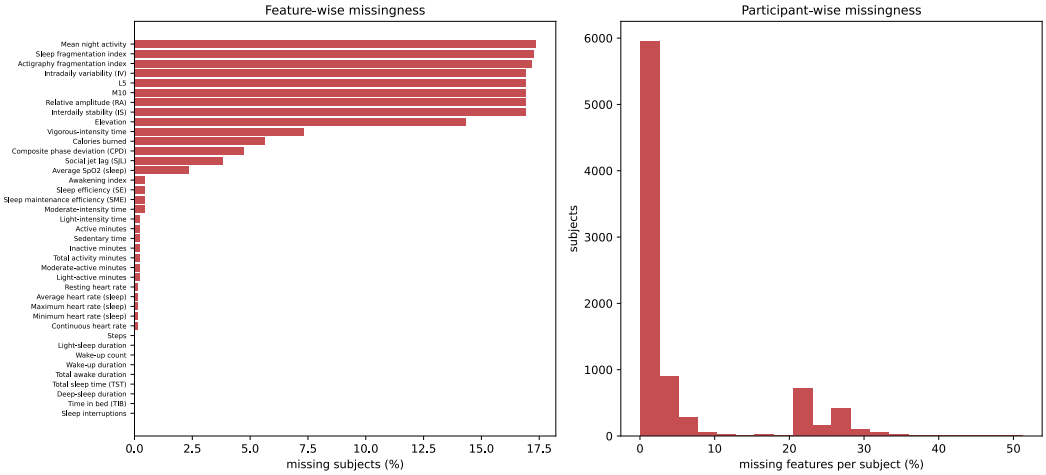


Figure D.5: Feature-wise and participant-wise missingness in the objective feature matrix before imputation.

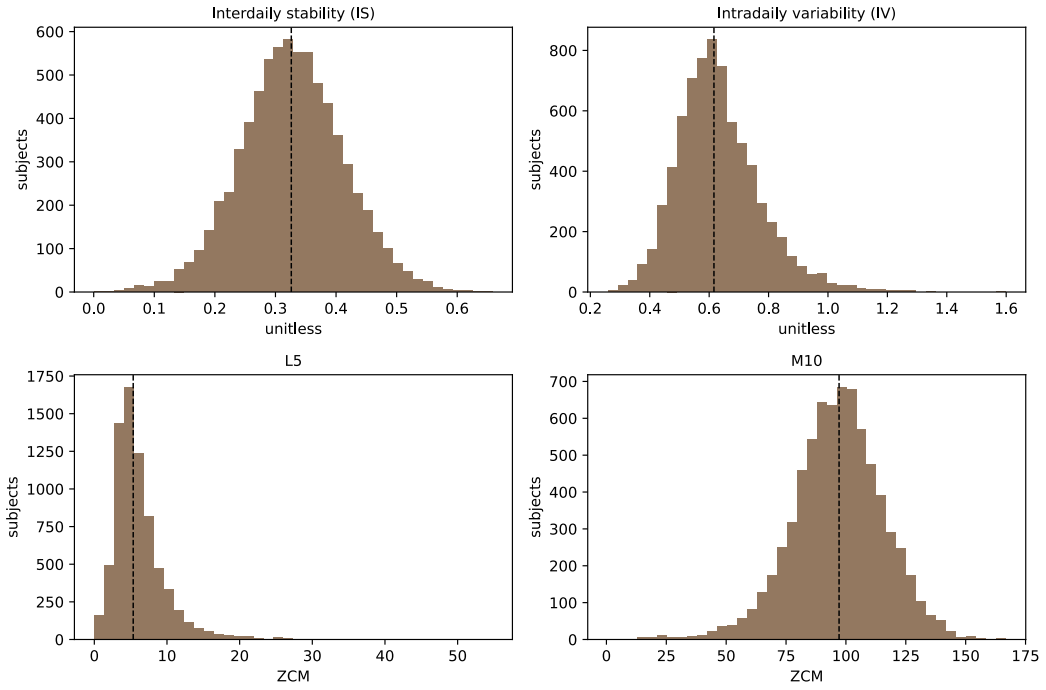


Figure D.6: Distributions of the nonparametric circadian metrics: interdaily stability, intradaily variability, L5 and M10.

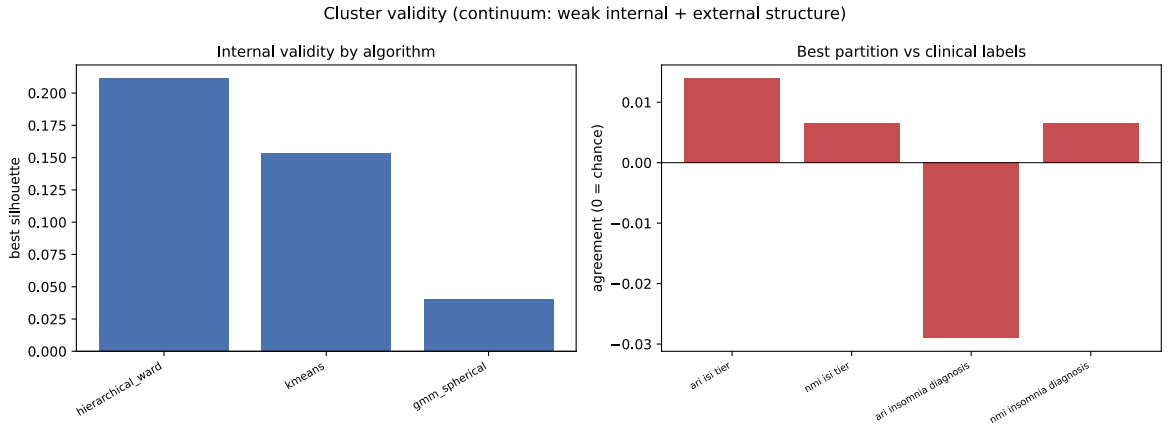


Figure D.7: Internal cluster-validity scores across the panel of algorithms and cluster counts. The best non-degenerate partition reaches a silhouette of only 0.21.

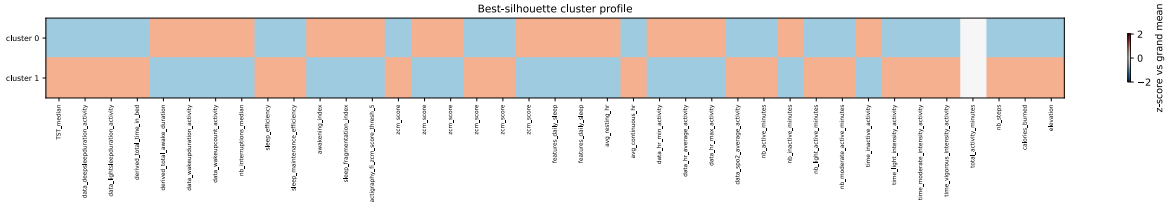


Figure D.8: Feature profile of the best available two-group partition (hierarchical, $k = 2$), shown to make explicit that the most favorable solution separates the cohort only weakly.

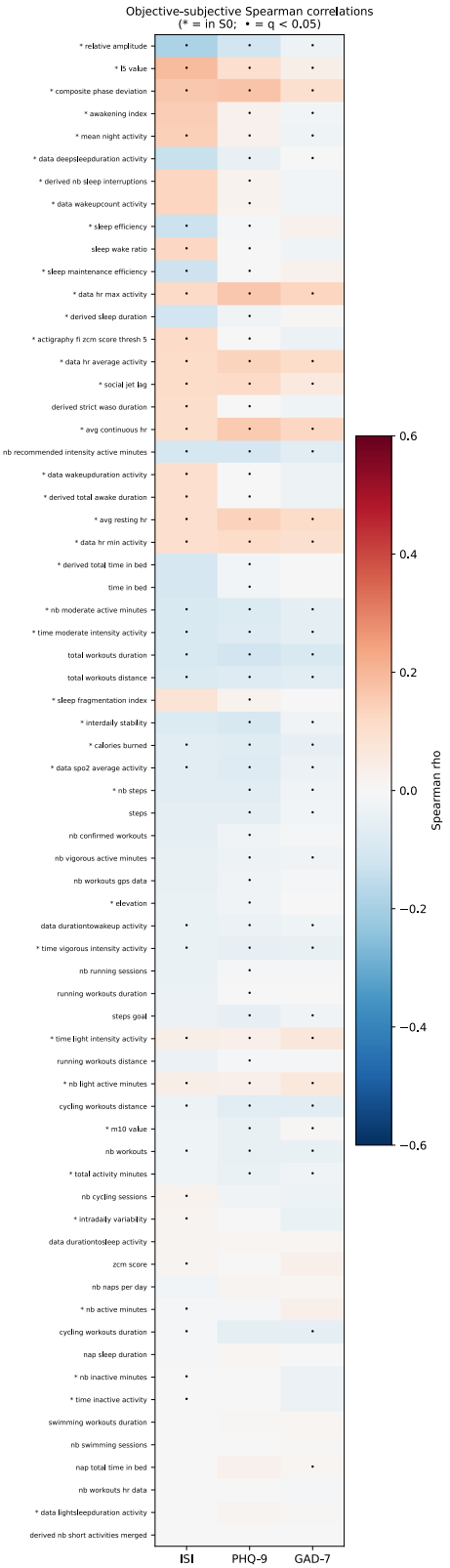


Figure D.9: Comprehensive screen: Spearman correlations of every objective feature with the ISI, PHQ-9 and GAD-7 scores, with Benjamini–Hochberg false-discovery control.

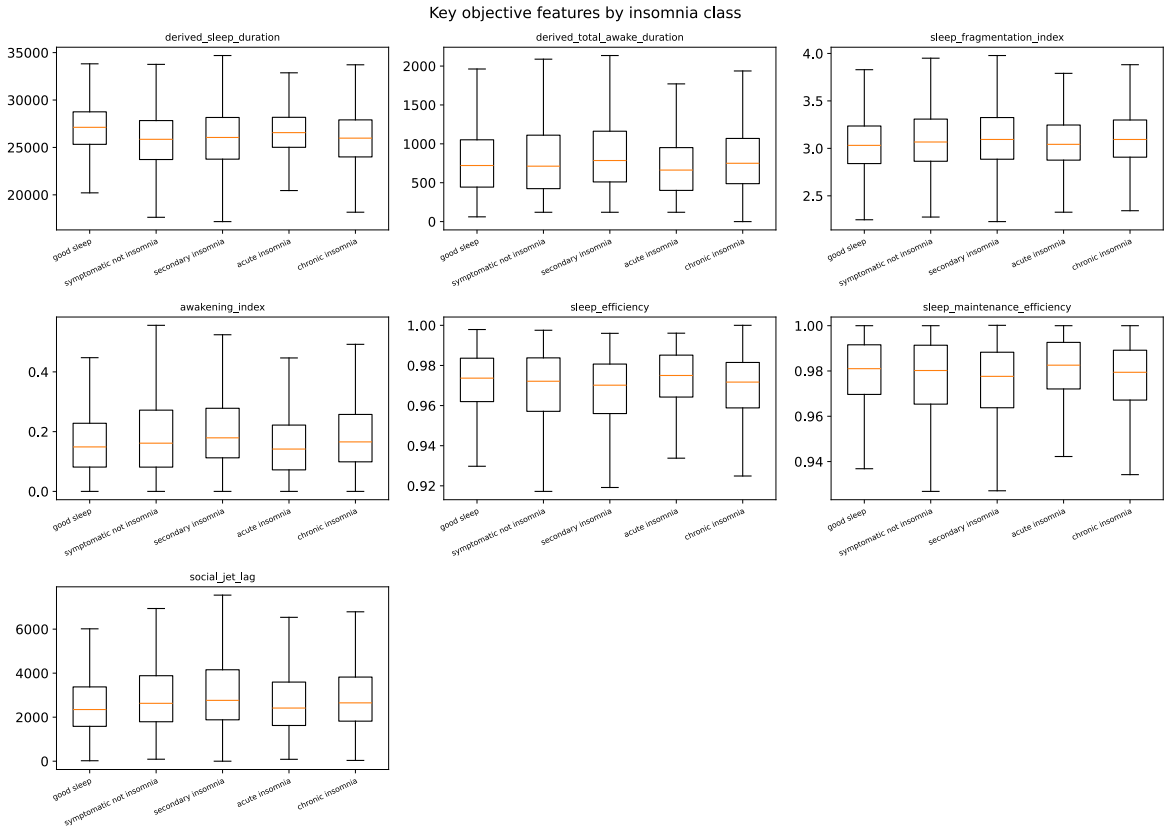


Figure D.10: Key objective sleep features across the DSM-5 insomnia classes.

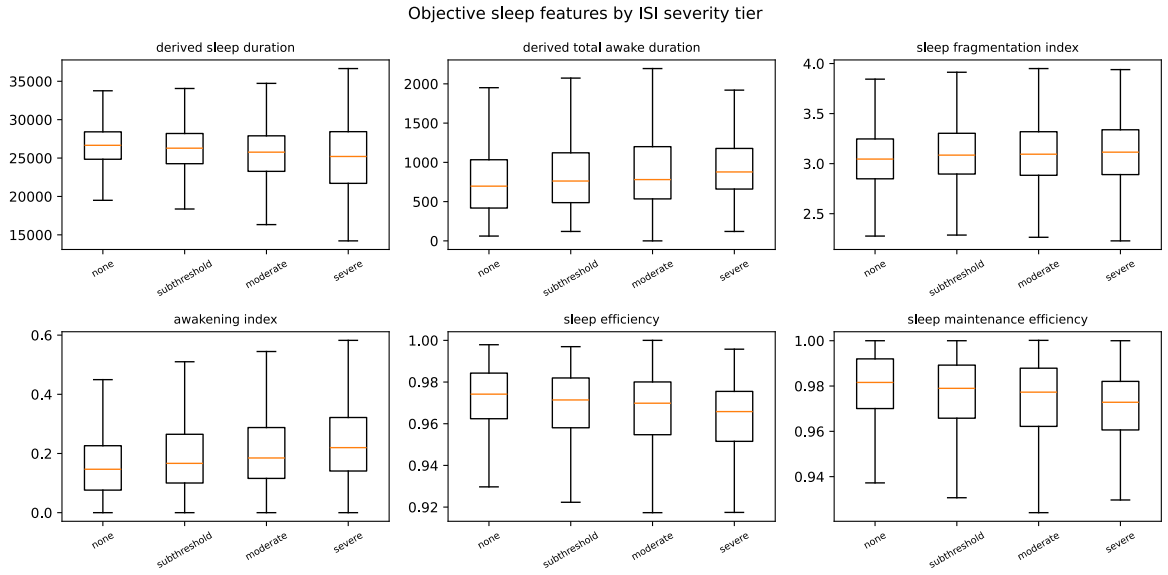


Figure D.11: Key objective sleep features across the ISI severity tiers.

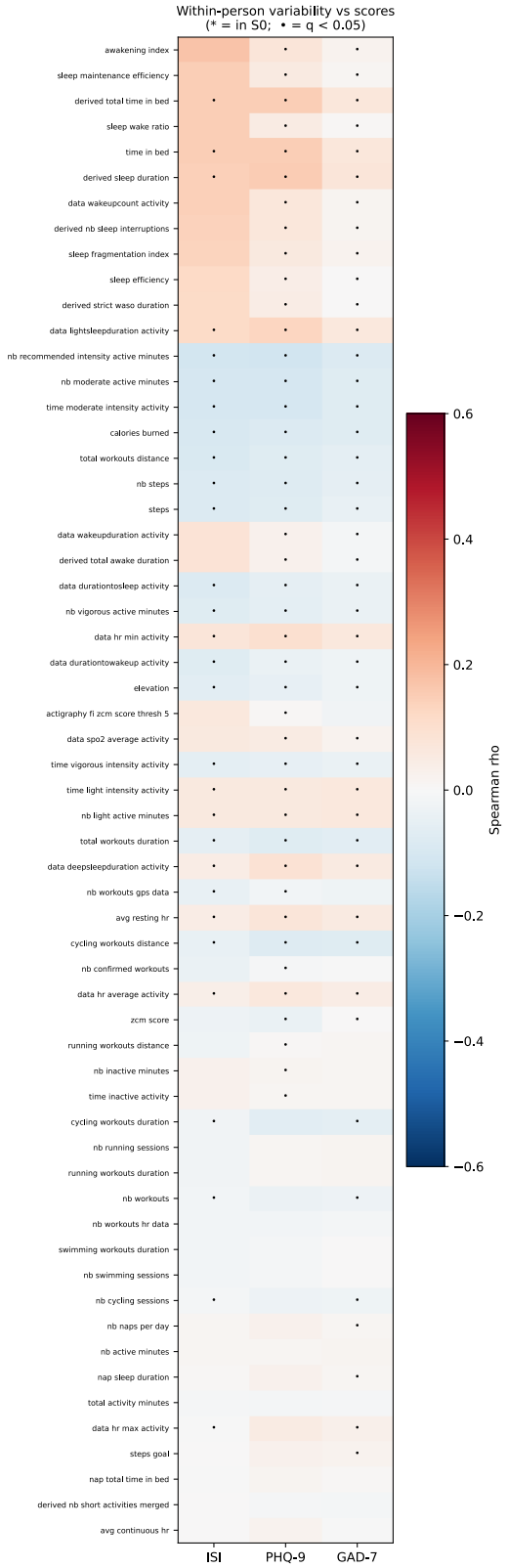


Figure D.12: Spearman correlations of the within-person standard-deviation features with the ISI, PHQ-9 and GAD-7 scores.

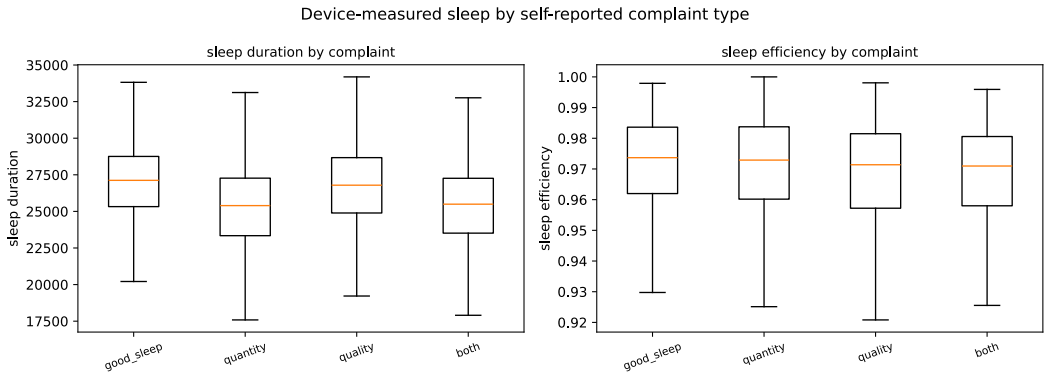


Figure D.13: Device-measured sleep duration and efficiency across the self-reported complaint dimension, a proxy for the unavailable sleep-state misperception comparison.

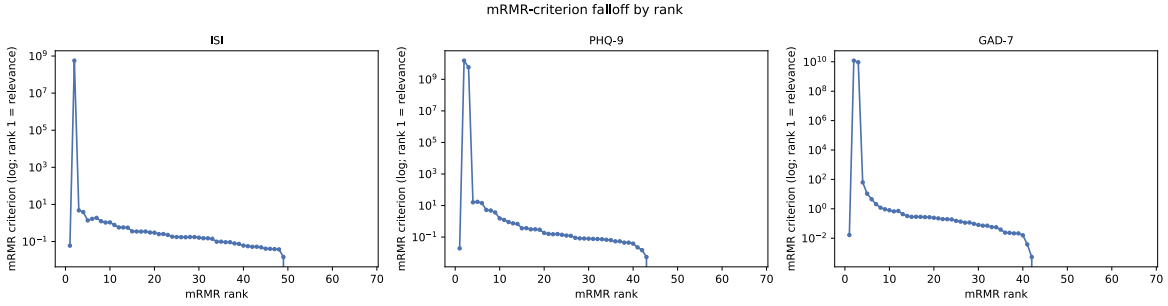


Figure D.14: Minimum-redundancy-maximum-relevance feature ranking per score: a descriptive ordering that favors features both relevant to the score and non-redundant with one another, and is not a validated selector since it uses the score to order them.

References

[1] Seung-Gul Kang et al. “Validity of a commercial wearable sleep tracker in adult insomnia disorder patients and good sleepers”. In: *Journal of psychosomatic research* 97 (2017), pp. 38–44.

[2] Bastien Lechat et al. “Multinight prevalence, variability, and diagnostic misclassification of obstructive sleep apnea”. In: *American journal of respiratory and critical care medicine* 205.5 (2022), pp. 563–569.

[3] Emilie Stern et al. “How can digital mental health enhance psychiatry?” In: *The Neuroscientist* 29.6 (2023), pp. 681–693.

[4] Jiaju Yin, Jiandong Xu, and Tian-Ling Ren. “Recent progress in long-term sleep monitoring technology”. In: *Biosensors* 13.3 (2023), p. 395.

[5] Adam V Benjafield et al. “Estimation of the global prevalence and burden of insomnia: a systematic literature review-based analysis”. In: *Sleep medicine reviews* 82 (2025), p. 102121.

[6] Damien Leger et al. “Prevalence of insomnia in a survey of 12 778 adults in France”. In: *Journal of sleep research* 9.1 (2000), pp. 35–42.

[7] American Psychiatric Association et al. *Diagnostic and statistical manual of mental disorders*. American psychiatric association, 2013.

[8] Isabelle Poirot et al. “Sommeil, somnolence et santé mentale–enquête INSV/Fondation VINCI Autoroutes pour la Journée du sommeil 2025”. In: *Médecine du Sommeil* 23.1 (2026), p. 69.

[9] Laura Palagini et al. “The need to prioritize ”Insomnia disorder” in public health agendas:“a wakeup call” position paper from European and Canadian experts in sleep and mental health”. In: *Sleep Medicine* (2025), p. 106763.

[10] *has-sante.fr*. https://www.has-sante.fr/upload/docs/application/pdf/2024-03/iqss_2024_guide_proms_obesite_2024.pdf. [Accessed 11-12-2025].

[11] Thomas Roth. “Insomnia: definition, prevalence, etiology, and consequences”. In: *Journal of clinical sleep medicine* 3.5 suppl (2007), S7–S10.

[12] Pierre A Geoffroy et al. “Insomnia and hypersomnia in major depressive episode: prevalence, sociodemographic characteristics and psychiatric comorbidity in a population-based study”. In: *Journal of affective disorders* 226 (2018), pp. 132–141.

[13] Damien Leger et al. “Insomnia and accidents: cross-sectional study (EQUINOX) on sleep-related home, work and car accidents in 5293 subjects with insomnia from 10 countries”. In: *Journal of sleep research* 23.2 (2014), pp. 143–152.

[14] Q. He et al. “The association between insomnia symptoms and risk of cardio-cerebral vascular events: A meta-analysis of prospective cohort studies”. In: *European Journal of Preventive Cardiology* 24.10 (2017), pp. 1071–1082. DOI: 10.1177/2047487317702043.

[15] S. C. Larsson and H. S. Markus. “Genetic Liability to Insomnia and Cardiovascular Disease Risk”. In: *Circulation* 140.9 (2019), pp. 796–798. DOI: 10.1161/CIRCULATIONAHA.119.041830.

[16] Alexandros N. Vgontzas et al. “Insomnia with objective short sleep duration: the most biologically severe phenotype of the disorder”. In: *Sleep Medicine Reviews* 17.4 (2013), pp. 241–254. DOI: 10.1016/j.smrv.2012.09.005.

[17] Yanyuan Dai et al. “Insomnia with objective, but not subjective, short sleep duration is associated with increased risk of incident hypertension: the Sleep Heart Health Study”. In: *Journal of Clinical Sleep Medicine* 19.8 (2023), pp. 1421–1428.

[18] Jesper Gottorp Riise et al. “Examining the burden of major depressive disorder with moderate-to-severe insomnia symptoms in five European countries”. In: *Journal of Affective Disorders* 386 (2025), p. 119403.

[19] Michael J Sateia. “International classification of sleep disorders”. In: *Chest* 146.5 (2014), pp. 1387–1394.

[20] David L Reed and William P Sacco. “Measuring sleep efficiency: what should the denominator be?” In: *Journal of clinical sleep medicine* 12.2 (2016), pp. 263–266.

[21] Laura S Castro et al. “Objective prevalence of insomnia in the São Paulo, Brazil epidemiologic sleep study”. In: *Annals of neurology* 74.4 (2013), pp. 537–546.

[22] Célyne H Bastien, Annie Vallières, and Charles M Morin. “Validation of the Insomnia Severity Index as an outcome measure for insomnia research”. In: *Sleep medicine* 2.4 (2001), pp. 297–307.

[23] DJ Buysse. “The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research”. In: *Psychiatry* 18 (2003), p. 498.

[24] Arthur J Spielman, Paul Saskin, and Michael J Thorpy. “Treatment of chronic insomnia by restriction of time in bed”. In: *Sleep* 10.1 (1987), pp. 45–56.

[25] Constantin R Soldatos, Dimitris G Dikeos, and Thomas J Paparrigopoulos. “Athens Insomnia Scale: validation of an instrument based on ICD-10 criteria”. In: *Journal of psychosomatic research* 48.6 (2000), pp. 555–560.

[26] Emilie Stern et al. “Validation of the French version of the Insomnia Severity Index in a population with psychiatric disorders and sleep complaints”. In: *L’Encéphale* (2025).

[27] Sleep Foundation. *Cognitive Behavioral Therapy for Insomnia (CBT-I): An Overview*. https://www.sleepfoundation.org/insomnia/treatment/cognitive-behavioral-therapy-insomnia. [Accessed 11-12-2025].

[28] Emilie Stern et al. “Treatment of insomnia disorder in psychiatry: A practical guide for CBT-I based intervention”. In: *L’encephale* (2026).

[29] D Leger and B Poursain. “An international survey of insomnia: under-recognition and under-treatment of a polysymptomatic condition”. In: *Current medical research and opinion* 21.11 (2005), pp. 1785–1792.

[30] M. M. Ohayon. “Epidemiology of insomnia: what we know and what we still need to learn”. In: *Sleep Medicine Reviews* 6.2 (2002), pp. 97–111. DOI: 10.1053/smrv.2002.0186.

[31] Dieter Riemann et al. “Insomnia disorder: State of the science and challenges for the future”. In: *Journal of Sleep Research* 31.4 (2022), e13604. DOI: 10.1111/jsr.13604.

[32] Jessica C. Levenson, Daniel B. Kay, and Daniel J. Buysse. “The pathophysiology of insomnia”. In: *Chest* 147.4 (2015), pp. 1179–1192. DOI: 10.1378/chest.14-1617.

[33] A. J. Spielman, L. S. Caruso, and P. B. Glovinsky. “A behavioral perspective on insomnia treatment”. In: *Psychiatric Clinics of North America* 10.4 (1987), pp. 541–553.

[34] Mackenzie J. Lind and Philip R. Gehrman. “Genetic Pathways to Insomnia”. In: *Brain Sciences* 6.4 (2016), p. 64. DOI: 10.3390/brainsci6040064.

[35] Pierre-Alexis Geoffroy et al. “Management of Insomnia Complaints by Non-Sleep Specialist Physicians: A French DELPHI Consensus”. In: *Journal of Sleep Research* 35.2 (2026), e70143.

[36] Dieter Riemann et al. “The European Insomnia Guideline: An update on the diagnosis and treatment of insomnia 2023”. In: *Journal of sleep research* 32.6 (2023), e14035.

[37] Alix Romier et al. “Objective sleep markers of suicidal behaviors in patients with psychiatric disorders: A systematic review and meta-analysis”. In: *Sleep medicine reviews* 68 (2023), p. 101760.

[38] Pierre Alexis Geoffroy et al. “Sleep in remitted bipolar disorder: a naturalistic case-control study using actigraphy”. In: *Journal of affective disorders* 158 (2014), pp. 1–7.

[39] Jessica Vensel Rundo and Ralph Downey III. “Polysomnography”. In: *Handbook of clinical neurology* 160 (2019), pp. 381–392.

[40] P. J. Hauri and E. M. Olmstead. “Reverse first night effect in insomnia”. In: *Sleep* 12.2 (1989), pp. 97–105. DOI: 10.1093/sleep/12.2.97.

[41] TeYang Lau et al. “Minimum number of nights for reliable estimation of habitual sleep using a consumer sleep tracker”. In: *Sleep Advances* 3.1 (2022), zpac026. DOI: 10.1093/sleepadvances/zpac026.

[42] Y. Zhang et al. “Worldwide and regional prevalence rates of co-occurrence of insomnia and insomnia symptoms with obstructive sleep apnea: A systematic review and meta-analysis”. In: *Sleep Medicine Reviews* 45 (2019), pp. 1–17. DOI: 10.1016/j.smrv.2019.01.004.

[43] A. M. Sweetman et al. “Developing a successful treatment for co-morbid insomnia and sleep apnoea”. In: *Sleep Medicine Reviews* 33 (2017), pp. 28–38. DOI: 10.1016/j.smrv.2016.04.004.

[44] A. Ghermezian et al. “A Predictive Model to Diagnose Psychophysiological Insomnia with Daytime Hyperarousal and Nighttime Micro–Macro-Structures Correlation”. English. In: *Sleep Vigil.* 7.2 (2023), pp. 123–133. ISSN: 25102265 (ISSN). DOI: 10.1007/s41782-023-00236-5. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85163792573&doi=10.1007%2fs41782-023-00236-5&partnerID=40&md5=005c45221454ac5b965656b856e7e2b0.

[45] M. H. Bonnet and D. L. Arand. “Activity, arousal, and the MSLT in patients with insomnia.” eng. In: *Sleep* 23.2 (Mar. 2000). Place: United States, pp. 205–212. ISSN: 0161-8105. DOI: 10.1093/sleep/23.2.1g.

[46] Florence M. Aellen et al. “Disentangling the complex landscape of sleep-wake disorders with data-driven phenotyping: A study of the Bernese center.” eng. In: *Eur J Neurol* 31.1 (Jan. 2024). Place: England, e16026. ISSN: 1468-1331 1351-5101. DOI: 10.1111/ene.16026.

[47] M Montserrat Sánchez-Ortuño et al. “Home Is Where Sleep Is: An Ecological Approach to Test the Validity of Actigraphy for the Assessment of Insomnia”. In: *Journal of Clinical Sleep Medicine* 6.1 (2010), pp. 21–29. DOI: 10.5664/jcsm.27706.

[48] Marie-Antoinette Spina et al. “Does changing perceptions of sleep by incorporating sleep wearables improve insomnia? Protocol for a randomized study (the Novel Insomnia Treatment Experiment)”. In: *Sleep Advances* 4.1 (2023), zpad012.

[49] Paul Edouard et al. “Validation of the Withings Sleep Analyzer, an under-the-mattress device for the detection of moderate-severe sleep apnea syndrome”. In: *Journal of Clinical Sleep Medicine* 17.6 (2021), pp. 1217–1227.

[50] Romain Kirszenblat and Paul Edouard. “Validation of the Withings ScanWatch as a wrist-worn reflective pulse oximeter: prospective interventional clinical study”. In: *Journal of Medical Internet Research* 23.4 (2021), e27503.

[51] Aymen A Alian and Kirk H Shelley. "Photoplethysmography". In: *Best Practice & Research Clinical Anaesthesiology* 28.4 (2014), pp. 395–406.

[52] Toshiyo Tamura. "Current progress of photoplethysmography and SPO2 for health monitoring". In: *Biomedical engineering letters* 9.1 (2019), pp. 21–36.

[53] Mostafa Shahin et al. "Deep Learning and Insomnia: Assisting Clinicians With Their Diagnosis". en. In: *IEEE J. Biomed. Health Inform.* 21.6 (Nov. 2017). 64, pp. 1546–1553. ISSN: 2168-2194, 2168-2208. DOI: 10.1109/JBHI.2017.2650199. URL: https://ieeexplore.ieee.org/document/7811237/ (visited on 05/11/2023).

[54] Mostafa Shahin et al. "A two stage approach for the automatic detection of insomnia". In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2018, pp. 466–469.

[55] Manish Sharma, Harsh S. Dhiman, and U. Rajendra Acharya. "Automatic identification of insomnia using optimal antisymmetric biorthogonal wavelet filter bank with ECG signals". en. In: *Computers in Biology and Medicine* 131 (Apr. 2021). 35, p. 104246. ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2021.104246. URL: https://www.sciencedirect.com/science/article/pii/S0010482521000408 (visited on 05/19/2023).

[56] Ruben Fossion et al. "Multiscale adaptive analysis of circadian rhythms and intradaily variability: Application to actigraphy time series in acute insomnia subjects". en. In: *PLOS ONE* 12.7 (2017). 46, e0181762. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0181762. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0181762 (visited on 05/19/2023).

[57] SS Kim. "Recent trends of artificial intelligence and machine learning for insomnia research". In: *Chronobiology in Medicine* 3.1 (2021), pp. 16–19.

[58] Dongmei He et al. "Insomnia disorder diagnosed by resting-state fMRI-based SVM classifier". In: *Sleep medicine* 95 (2022), pp. 126–129.

[59] Maia Angelova et al. "Automated Method for Detecting Acute Insomnia Using Multi-Night Actigraphy Data". In: *IEEE ACCESS* 8 (2020), pp. 74413–74422. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.2988722.

[60] Sungkyu Park et al. "Clustering Insomnia Patterns by Data From Wearable Devices: Algorithm Development and Validation Study". en. In: *JMIR Mhealth Uhealth* 7.12 (Dec. 2019). 11, e14473. ISSN: 2291-5222. DOI: 10.2196/14473. URL: https://mhealth.jmir.org/2019/12/e14473 (visited on 05/23/2023).

[61] Sleep Foundation. *Stages of Sleep: What Happens in a Normal Sleep Cycle?* https://www.sleepfoundation.org/stages-of-sleep. [Accessed 11-12-2025].

[62] Pragati Tripathi et al. “Ensemble computational intelligent for insomnia sleep stage detection via the sleep ECG signal”. In: *IEEE Access* 10 (2022), pp. 108710–108721.

[63] A Goldberger et al. “CAP Sleep Database”. In: ().

[64] Sifan Hu et al. “First-night effect in insomnia disorder: a systematic review and meta-analysis of polysomnographic findings”. In: *Journal of sleep research* 33.1 (2024), e13942.

[65] Y Cha. “Practical artificial intelligence for physicians”. In: *Paju: Koonja* 6 (2017), pp. 88–93.

[66] Sirvan Khalighi et al. “ISRUC-Sleep: A comprehensive public dataset for sleep researchers”. In: *Computer methods and programs in biomedicine* 124 (2016), pp. 180–192.

[67] P. Wasiewicz, M. Skalski, and M. Fornal-Pawlowska. “Chronic insomnia cases detection with a help of Athens Insomnia Scale and SF-36 health survey”. In: *PHOTONICS APPLICATIONS IN ASTRONOMY, COMMUNICATIONS, INDUSTRY, AND HIGH-ENERGY PHYSICS EXPERIMENTS 2011*. Ed. by RS Romaniuk. Vol. 8008. Proceedings of SPIE. Inst Elect Syst; Fac Elect & Informat Technol; Warsaw Univ Technol; Photon Soc Poland (PSP); Polish Acad Sci, Comm Elect & Telecommun; European Coordinat Accelerator R&D (EU FP7) (EuCARD); IEEE Poland Sect; Polish Comm Optoelect SEP (PKOpto), 2011. ISBN: 978-0-8194-8582-3. DOI: 10.1117/12.905587.

[68] Wei Qu et al. “Single-channel EEG based insomnia detection with domain adaptation”. en. In: *Computers in Biology and Medicine* 139 (Dec. 2021). 3, p. 104989. ISSN: 00104825. DOI: 10.1016/j.compbiomed.2021.104989. URL: https://linkinghub.elsevier.com/retrieve/pii/S0010482521007836 (visited on 05/11/2023).

[69] Annemarie I Luik, Simon D Kyle, and Colin A Espie. “Digital cognitive behavioral therapy (dCBT) for insomnia: a state-of-the-science review”. In: *Current sleep medicine reports* 3.2 (2017), pp. 48–56.

[70] Christophe Gauld et al. “A systematic review of sleep–wake disorder diagnostic criteria reliability studies”. In: *Biomedicines* 10.7 (2022), p. 1616.

[71] Scott Ryals et al. “Photoplethysmography—new applications for an old technology: a sleep technology review”. In: *Journal of Clinical Sleep Medicine* 19.1 (2023), pp. 189–195.

[72] Shuting Xu et al. “A review of automated sleep disorder detection”. In: *Computers in Biology and Medicine* 150 (2022), p. 106100. DOI: 10.1016/j.compbiomed.2022.106100.

[73] Manisha Ingle et al. “A systematic review on automatic identification of insomnia”. In: *Physiological Measurement* 45.3 (2024), 03TR01. DOI: 10.1088/1361-6579/ad2059.

[74] Md Belal Bin Heyat et al. "Progress in Detection of Insomnia Sleep Disorder: A Comprehensive Review". In: *Current Drug Targets* 22.6 (2021), pp. 672–684. DOI: 10.2174/1389450121666201027125828.

[75] Hannah Scott et al. "Emerging applications of objective sleep assessments towards the improved management of insomnia". In: *Sleep Medicine* 101 (2023), pp. 138–145. DOI: 10.1016/j.sleep.2022.10.030.

[76] Matthew J Page et al. "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews". In: *bmj* 372 (2021).

[77] *Zotero | Your personal research assistant — zotero.org*. https://www.zotero.org/. [Accessed 20-01-2026].

[78] *Merge all duplicates — forums.zotero.org*. https://forums.zotero.org/discussion/40457/merge-all-duplicates#:~:text=var%20DupPane%20%3D%20Zotero.getZoteroPanes()%3B%0Afor(var%20i%20%3D%200%3B%20i%20%3C%20100%3B%20i%2B%2B)%20%7B%0Aawait%20new%20Promise(r%20%3D%3E%20setTimeout(r%2C%201000))%3B%0ADupPane%5B0%5D.mergeSelectedItems()%3B%0AZotero_Duplicates_Pane.merge()%3B%0A%7D. [Accessed 20-01-2026].

[79] Kathy L Nelson, Jean E Davis, and Cynthia F Corbett. "Sleep quality: An evolutionary concept analysis". In: *Nursing forum*. Vol. 57. 1. Wiley Online Library. 2022, pp. 144–151.

[80] A. Wadichar et al. "A Hierarchical Approach for the Diagnosis of Sleep Disorders Using Convolutional Recurrent Neural Network". English. In: *IEEE Access* 11 (2023), pp. 125244–125255. ISSN: 21693536 (ISSN). DOI: 10.1109/ACCESS.2023.3330901. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85177064092&doi=10.1109%2fACCESS.2023.3330901&partnerID=40&md5=4c029d15c44f7d8df5552c05370d01c8.

[81] Manish Sharma, Virendra Patel, and U. Rajendra Acharya. "Automated identification of insomnia using optimal bi-orthogonal wavelet transform technique with single-channel EEG signals". In: *KNOWLEDGE-BASED SYSTEMS* 224 (July 2021). ISSN: 0950-7051. DOI: 10.1016/j.knosys.2021.107078.

[82] Manish Sharma et al. "Automated Identification of Sleep Disorder Types Using Triplet Half-Band Filter and Ensemble Machine Learning Techniques with EEG Signals". In: *ELECTRONICS* 10.13 (July 2021). DOI: 10.3390/electronics10131531.

[83] Manish Sharma et al. "Automated identification of sleep disorders using wavelet-based features extracted from electrooculogram and electromyogram signals." eng. In: *Comput Biol Med* 143 (Jan. 2022). Place: United States, p. 105224. ISSN: 1879-0534 0010-4825. DOI: 10.1016/j.compbiomed.2022.105224.

[84] M. Sharma et al. “Automated insomnia detection using wavelet scattering network technique with single-channel EEG signals”. English. In: *Eng Appl Artif Intell* 126 (2023). ISSN: 09521976 (ISSN). DOI: 10.1016/j.engappai.2023.106903. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85170437732&doi=10.1016%2fj.engappai.2023.106903&partnerID=40&md5=45db99deea2ea0fcc355c481b90f6655.

[85] Manish Sharma et al. “Sleep disorder identification using wavelet scattering on ECG signals”. en. In: *International Journal of Imaging Systems and Technology* n/a.n/a (2024). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ima.22980. ISSN: 1098-1098. DOI: 10.1002/ima.22980. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ima.22980 (visited on 11/19/2023).

[86] Y.-H. Cheng, M. Lech, and R.H. Wilkinson. “Simultaneous Sleep Stage and Sleep Disorder Detection from Multimodal Sensors Using Deep Learning”. In: *Sensors* 23.7 (2023). DOI: 10.3390/s23073468. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85152318612&doi=10.3390%2fs23073468&partnerID=40&md5=dc29e673318965ebcb5378dc60ee3f6a.

[87] S. Rani et al. “Differentiating acute from chronic insomnia with machine learning from actigraphy time series data.” eng. In: *Front Netw Physiol* 2 (2022). Place: Switzerland, p. 1036832. ISSN: 2674-0109. DOI: 10.3389/fnetp.2022.1036832.

[88] Stephen McCloskey et al. “Insomnia Disorder Detection using EEG Sleep Trajectories”. In: *ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING, PAKDD 2022, PT III*. Ed. by J Gama et al. Vol. 13282. Lecture Notes in Artificial Intelligence. 2022, pp. 314–325. ISBN: 978-3-031-05981-0 978-3-031-05980-3. DOI: 10.1007/978-3-031-05981-0_25.

[89] Stephen McCloskey et al. “Data-driven cluster analysis of insomnia disorder with physiology-based qEEG variables”. In: *KNOWLEDGE-BASED SYSTEMS* 183 (Nov. 2019). ISSN: 0950-7051. DOI: 10.1016/j.knosys.2019.07.034.

[90] Beatriz Rodriguez-Morilla et al. “Application of Machine Learning Methods to Ambulatory Circadian Monitoring (ACM) for Discriminating Sleep and Circadian Disorders.” eng. In: *Front Neurosci* 13 (2019). Place: Switzerland, p. 1318. ISSN: 1662-4548 1662-453X. DOI: 10.3389/fnins.2019.01318.

[91] Mauro Manconi et al. “Measuring the error in sleep estimation in normal subjects and in patients with insomnia.” eng. In: *J Sleep Res* 19.3 (Sept. 2010). Place: England, pp. 478–486. ISSN: 1365-2869 0962-1105. DOI: 10.1111/j.1365-2869.2009.00801.x.

[92] Mi Hyun Lee et al. "Multitask fMRI and machine learning approach improve prediction of differential brain activity pattern in patients with insomnia disorder." eng. In: *Sci Rep* 11.1 (Apr. 2021). Place: England, p. 9402. ISSN: 2045-2322. DOI: 10.1038/s41598-021-88845-w.

[93] M. Warunlawan et al. "Identification of Crucial Factors in Sleep Quality Using Machine Learning Models and MRMR Feature Selection Technique". In: 2023. DOI: 10.1109/BMEiCON60347.2023.10322011. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85179550923&doi=10.1109%2fBMEiCON60347.2023.10322011&partnerID=40&md5=75968386503d9c347b8b3376e67dd08c.

[94] Robert Joseph Thomas, Christopher Wood, and Matt Travis Bianchi. "Cardiopulmonary coupling spectrogram as an ambulatory clinical biomarker of sleep stability and quality in health, sleep apnea, and insomnia." eng. In: *Sleep* 41.2 (Feb. 2018). Place: United States. ISSN: 1550-9109 0161-8105. DOI: 10.1093/sleep/zsx196.

[95] M. Afshani et al. "Discriminating Paradoxical and Psychophysiological Insomnia Based on Structural and Functional Brain Images: A Preliminary Machine Learning Study". In: *Brain Sciences* 13.4 (2023). DOI: 10.3390/brainsci13040672. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85156212968&doi=10.3390%2fbrainsci13040672&partnerID=40&md5=a1e9888f1776eefedf1dcb22ebb5ae4a.

[96] Yishul Wei et al. "EEG Microstates Indicate Heightened Somatic Awareness in Insomnia: Toward Objective Assessment of Subjective Mental Content." eng. In: *Front Psychiatry* 9 (2018). Place: Switzerland, p. 395. ISSN: 1664-0640. DOI: 10.3389/fpsyt.2018.00395.

[97] Peter T. Morgan et al. "Cortical GABA Levels in Primary Insomnia". In: *SLEEP* 35.6 (June 2012), pp. 807–814. ISSN: 0161-8105. DOI: 10.5665/sleep.1880.

[98] Christopher B. Miller et al. "Clusters of Insomnia Disorder: An Exploratory Cluster Analysis of Objective Sleep Parameters Reveals Differences in Neurocognitive Functioning, Quantitative EEG, and Heart Rate Variability." eng. In: *Sleep* 39.11 (Nov. 2016). Place: United States, pp. 1993–2004. ISSN: 1550-9109 0161-8105. DOI: 10.5665/sleep.6230.

[99] S. Kusmakar et al. "A machine learning model for multi-night actigraphic detection of chronic insomnia: development and validation of a pre-screening tool". en. In: *R. Soc. open sci.* 8.6 (June 2021). 7, p. 202264. ISSN: 2054-5703. DOI: 10.1098/rsos.202264. URL: https://royalsocietypublishing.org/doi/10.1098/rsos.202264 (visited on 05/11/2023).

[100] Chi Zhang et al. “Quantitative evaluation of short-term resting-state brain networks for primary insomnia diagnosis”. In: *BIOMEDICAL SIGNAL PROCESSING AND CONTROL* 74 (Apr. 2022). ISSN: 1746-8094. DOI: 10.1016/j.bspc.2022.103498.

[101] Lorenzo Tonetti et al. “Advantages of single step over step-by-step screening for sleep disorders”. In: *BIOLOGICAL RHYTHM RESEARCH* 49.4 (2018), pp. 610–621. ISSN: 0929-1016. DOI: 10.1080/09291016.2017.1390822.

[102] J. D. Edinger et al. “The empirical identification of insomnia subtypes: a cluster analytic approach.” eng. In: *Sleep* 19.5 (June 1996). Place: United States, pp. 398–411. ISSN: 0161-8105. DOI: 10.1093/sleep/19.5.398.

[103] Thomas Andrillon et al. “Revisiting the value of polysomnographic data in insomnia: more than meets the eye”. en. In: *Sleep Medicine* 66 (Feb. 2020), pp. 184–200. ISSN: 1389-9457. DOI: 10.1016/j.sleep.2019.12.002. URL: https://www.sciencedirect.com/science/article/pii/S1389945719316442 (visited on 05/24/2023).

[104] Cheng-Yu Tsai et al. “Differentiation Model for Insomnia Disorder and the Respiratory Arousal Threshold Phenotype in Obstructive Sleep Apnea in the Taiwanese Population Based on Oximetry and Anthropometric Features.” eng. In: *Diagnostics (Basel)* 12.1 (Dec. 2021). Place: Switzerland. ISSN: 2075-4418. DOI: 10.3390/diagnostics12010050.

[105] Qian Lu et al. “Connectomic disturbances underlying insomnia disorder and predictors of treatment response”. In: *FRONTIERS IN HUMAN NEUROSCIENCE* 16 (Aug. 2022). ISSN: 1662-5161. DOI: 10.3389/fnhum.2022.960350.

[106] Megan R. Crawford et al. “Characterization of Patients Who Present With Insomnia: Is There Room for a Symptom Cluster-Based Approach?” eng. In: *J Clin Sleep Med* 13.7 (July 2017). Place: United States, pp. 911–921. ISSN: 1550-9397 1550-9389. DOI: 10.5664/jcsm.6666.

[107] Soomi Lee, Brian D. Gonzalez, and Brent J. Small. “My job impacts my sleep: signs and symptoms of insomnia among healthcare workers.” eng. In: *Ind Health* 59.2 (2021). Place: Japan, pp. 86–98. ISSN: 1880-8026 0019-8366. DOI: 10.2486/indhealth.2020-0191.

[108] Chen Lin et al. “Examining Human-Smartphone Interaction as a Proxy for Circadian Rhythm in Patients With Insomnia: Cross-Sectional Study.” eng. In: *J Med Internet Res* 25 (Dec. 2023). Place: Canada, e48044. ISSN: 1438-8871 1439-4456. DOI: 10.2196/48044.

[109] P. J. Hauri. “A cluster analysis of insomnia.” eng. In: *Sleep* 6.4 (1983). Place: United States, pp. 326–338. ISSN: 0161-8105. DOI: 10.1093/sleep/6.4.326.

[110] Bart H. W. Te Lindert et al. "Actigraphic multi-night home-recorded sleep estimates reveal three types of sleep misperception in Insomnia Disorder and good sleepers." eng. In: *J Sleep Res* 29.1 (Feb. 2020). Place: England, e12937. ISSN: 1365-2869 0962-1105. DOI: 10.1111/jsr.12937.

[111] P. Wasiewicz and M. Skalski. "On insomnia analysis using methods of artificial intelligence". In: *PHOTONICS APPLICATIONS IN ASTRONOMY, COMMUNICATIONS, INDUSTRY, AND HIGH-ENERGY PHYSICS EXPERIMENTS 2011*. Ed. by RS Romaniuk. Vol. 8008. Proceedings of SPIE. Inst Elect Syst; Fac Elect & Informat Technol; Warsaw Univ Technol; Photon Soc Poland (PSP); Polish Acad Sci, Comm Elect & Telecommun; European Coordinat Accelerator R&D (EU FP7) (EuCARD); IEEE Poland Sect; Polish Comm Optoelect SEP (PKOpto), 2011. ISBN: 978-0-8194-8582-3. DOI: 10.1117/12.905590.

[112] Umaer Hanif et al. "Automatic Detection of Chronic Insomnia from Polysomnographic and Clinical Variables Using Machine Learning." eng. In: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*. Vol. 2023. Place: United States. July 2023, pp. 1–5. DOI: 10.1109/EMBC40787.2023.10340587.

[113] Vincent Mysliwiec et al. "Comorbid insomnia and obstructive sleep apnea in military personnel: correlation with polysomnographic variables." eng. In: *Mil Med* 179.3 (Mar. 2014). Place: England, pp. 294–300. ISSN: 1930-613X 0026-4075. DOI: 10.7205/MILMED-D-13-00396.

[114] Meredith L. Wallace et al. "Variable Selection for Skewed Model-Based Clustering: Application to the Identification of Novel Sleep Phenotypes." eng. In: *J Am Stat Assoc* 113.521 (2018). Place: United States, pp. 95–110. ISSN: 0162-1459 1537-274X. DOI: 10.1080/01621459.2017.1330202.

[115] Haslaile Abdullah et al. "Support Vector Machine Classification of EEG Nonlinear Features for Primary Insomnia". In: *2ND INTERNATIONAL CONFERENCE FOR INNOVATION IN BIOMEDICAL ENGINEERING AND LIFE SCIENCES*. Ed. by F Ibrahim et al. Vol. 67. IFMBE Proceedings. 2018, pp. 161–164. ISBN: 978-981-10-7554-4 978-981-10-7553-7. DOI: 10.1007/978-981-10-7554-4_28.

[116] Maia Angelova, Philip Holloway, and Jason Ellis. "Complexity analysis of human physiological signals based on case studies". In: *XXXTH INTERNATIONAL COLLOQUIUM ON GROUP THEORETICAL METHODS IN PHYSICS (ICGTMP) (GROUP30)*. Journal of Physics Conference Series 597 (2015). ISSN: 1742-6588. DOI: 10.1088/1742-6596/597/1/012010.

[117] L. Thiesse et al. "Somno-Art Software identifies pathology-induced changes in sleep parameters similarly to polysomnography". English. In: *PLoS ONE* 18.10 October (2023). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0291593.

[118] Sana Tmar-Ben Hamida, Beena Ahmed, and Thomas Penzel. “A Novel Insomnia Identification Method Based on Hjorth Parameters”. In: *2015 IEEE INTERNATIONAL SYMPOSIUM ON SIGNAL PROCESSING AND INFORMATION TECHNOLOGY (ISSPIT)*. IEEE International Symposium on Signal Processing and Information Technology. 2015, pp. 548–552. ISBN: 978-1-5090-0480-5. DOI: 10.1109/isspit.2015.7394397.

[119] Jacqueline Chandler et al. “Cochrane handbook for systematic reviews of interventions”. In: *Hoboken: Wiley* 4.1002 (2019), p. 14651858.

[120] S. Suh et al. “Clinical significance of night-to-night sleep variability in insomnia”. In: *Sleep Medicine* 13.5 (2012), pp. 469–475. DOI: 10.1016/j.sleep.2011.10.034.

[121] Massimiliano de Zambotti et al. “Wearable Sleep Technology in Clinical and Research Settings”. In: *Medicine & Science in Sports & Exercise* 51.7 (2019), pp. 1538–1557. DOI: 10.1249/MSS.0000000000001947.

[122] M.-A. Stefanos et al. “Evaluation of a Contactless Sleep Monitoring Device for Sleep Stage Detection at Home in a Healthy Population: Prospective Study in Free-Living Conditions”. In: *JMIR Human Factors* 13 (2026), e77033. DOI: 10.2196/77033.

[123] Y. J. Lee et al. “Interrater reliability of sleep stage scoring: a meta-analysis”. In: *Journal of Clinical Sleep Medicine* 18.1 (2022), pp. 193–202. DOI: 10.5664/jcsm.9538.

[124] S. Nikkonen et al. “Multicentre sleep-stage scoring agreement in the Sleep Revolution project”. In: *Journal of Sleep Research* 33.1 (2024), e13956. DOI: 10.1111/jsr.13956.

[125] Luca Menghini et al. “A standardized framework for testing the performance of sleep-tracking technology: step-by-step guidelines and open-source code”. In: *Sleep* 44.2 (2021), zsaa170. DOI: 10.1093/sleep/zsaa170.

[126] Leeba Rezaie et al. “Paradoxical insomnia and subjective-objective sleep discrepancy: A review”. In: *Sleep Medicine Reviews* 40 (2018), pp. 196–202. DOI: 10.1016/j.smrv.2018.01.002.

[127] Bei Bei et al. “Beyond the mean: A systematic review on the correlates of daily intraindividual variability of sleep/wake patterns”. In: *Sleep Medicine Reviews* 28 (2016), pp. 104–120. DOI: 10.1016/j.smrv.2015.06.003.

[128] T. F. Blanken et al. “Insomnia disorder subtypes derived from life history and traits of affect and personality”. In: *The Lancet Psychiatry* 6.2 (2019), pp. 151–163. DOI: 10.1016/S2215-0366(18)30464-4.

[129] Jeroen S. Benjamins et al. “Insomnia heterogeneity: Characteristics to consider for data-driven multivariate subtyping”. In: *Sleep Medicine Reviews* 36 (2017), pp. 71–81. DOI: 10.1016/j.smrv.2016.10.005.

[130] M. Carpi, D. R. Marques, and C. Liguori. "Deriving Subtypes From the Insomnia Severity Index: A Latent Class Analysis and Comparison of Features". In: *Behavioral Sleep Medicine* 23.6 (2025), pp. 807–819. DOI: 10.1080/15402002.2025.2539961.

[131] Charles M. Morin et al. "The natural history of insomnia: a population-based 3-year longitudinal study". In: *Archives of Internal Medicine* 169.5 (2009), pp. 447–453. DOI: 10.1001/archinternmed.2008.610.

[132] Charles M. Morin et al. "Incidence, Persistence, and Remission Rates of Insomnia Over 5 Years". In: *JAMA Network Open* 3.11 (2020), e2018782. DOI: 10.1001/jamanetworkopen.2020.18782.

[133] Yatharth Ranjan et al. "RADAR-Base: Open Source Mobile Health Platform for Collecting, Monitoring, and Analyzing Data Using Sensors, Wearables, and Mobile Devices". In: *JMIR mHealth and uHealth* 7.8 (2019), e11734. DOI: 10.2196/11734.

[134] Kelton Minor et al. "Rising temperatures erode human sleep globally". In: *One Earth* 5.5 (2022), pp. 534–549. DOI: 10.1016/j.oneear.2022.04.008.