

Causal Inference-Based Sleep Stage Scoring under Cluster-level Unobserved Confounding

Yeongho Lee*
yeongho.lee@thyroscope.com
THYROSCOPE INC.
Ulsan, Republic of Korea

Kyubo Shin
kyubo.shin@thyroscope.com
THYROSCOPE INC.
Ulsan, Republic of Korea

W. Hong Yeo†
whyeo@gatech.edu
Georgia Institute of Technology
George W. Woodruff School of Mechanical Engineering
Wearable Intelligent Systems and Healthcare Center,
Institute for Matter and Systems
Wallace H. Coulter Department of Biomedical Engineering
Parker H. Petit Institute for Bioengineering and
Biosciences
Atlanta, GA, USA

Gi-Soo Kim†
gisookim@unist.ac.kr
UNIST
Artificial Intelligence Graduate School
Department of Industrial Engineering
Ulsan, Republic of Korea

Abstract

Standard supervised machine learning models often fail on real-world biosignal data by learning spurious correlations from confounders (e.g., user or device variations), leading to poor out-of-distribution (OOD) generalization and unreliable interpretations. To address this, we introduce a causal inference framework that uses a novel cluster-adjusted Entropy Balancing for Continuous Treatments (EBCT) to reweight the data, decorrelating a key EEG feature (our Feature of Interest) from both observed confounders and unobserved hierarchical confounders. The classifiers trained on the reweighted data demonstrate robust OOD performance while unlocking causal interpretability through Dose-Response Functions (DRFs) that reveal the true effect of the EEG feature on sleep stages. This work presents a modular blueprint for building robust and causally-informed models from complex, hierarchical data.

Keywords

Causal Inference, Unobserved Confounding, Entropy Balancing, Interpretability, Robustness, Cluster adjustment, Sleep Staging, EEG

1 Introduction

High-quality sleep is increasingly recognized as a pillar of health, with broad links to cognition, metabolic and cardiovascular outcomes, and emotional regulation. In the clinical setting, polysomnography (PSG) remains the gold standard for sleep stage scoring into one of 5 stages—Wake, N1-N3, and REM—according to the American Academy of Sleep Medicine (AASM) rules [1, 19]. Despite its status as the gold standard, PSG presents significant hurdles for widespread use. The reliance on in-lab equipment and the laborious, expert-drive manual scoring process restricts its accessibility and scalability for routine or longitudinal monitoring. These practical

hurdles of PSG have therefore motivated a clear need for alternative approaches capable of providing accessible, scalable, and longitudinal sleep monitoring outside of the clinical setting.

Meanwhile, a rapid expansion of wearable sensor technology, has enabled rapid collection of massive biosignal data. The resulting data landscape has coincided with rapid improvements in machine learning, promising convenient, long-horizon sleep tracking outside the clinic [2, 8, 10, 24]. Remaining concerns involve the validity, generalizability, and downstream use of the automatic sleep scoring system trained on the massive but poorly curated data collected from wearables [2, 24].

Despite notable advances in sleep score modeling, models trained in one context often generalize poorly to others, especially when heterogeneity between contexts is high. Such heterogeneity can be induced by the use of different sensing devices, acquisition protocols, scoring conventions and the difference between the training and deployment populations [18]. The challenge sharpens when aforementioned factors act as *confounders*, i.e., affect both the signal features and the sleep score, hindering accurate modeling of their true relationship and creating spurious correlations. Such spurious correlations obstruct both generalization to heterogenous environment and a correct interpretation of the feature-sleep stage relationship.

We address these issues with a causal weighting approach. Specifically, we build on Entropy Balancing for Continuous Treatments (EBCT) [26], a convex reweighting scheme that enforces zero covariance between (continuous) feature of interest and other potential confounding factors [5, 28]. We denote the feature of interest as FOI and seek to estimate its true relationship to the sleep stage. Biosignal data and their associated sleep scores are clustered by subject IDs, which are in turn clustered by wearable devices they use. Often, the cluster-level unobserved factors can act as confounders. We introduce a cluster-adjusted EBCT that augments the original i.i.d. formulation with a within-cluster sign-balance

*First author.

†Corresponding author.

© 2025 Copyright held by the owner/author(s).

constraint on the standardized FOI, attenuating unobserved cluster-level confounding. In the resulting balanced pseudo-population, we perform weighted regression of sleep stage onto the FOI and other variables to (i) train robust sleep-staging system and (ii) estimate dose–response functions (DRFs) that quantify how stage probabilities vary with the intensity of FOI, when all other variables are fixed.

2 Related Work

2.1 Automatic Sleep Staging

Deep models for automatic sleep staging have evolved from early CNN–RNN hybrids to attention and transformer architectures that capture long-range temporal context and support interpretability. **DeepSleepNet** established a strong baseline by coupling convolutional feature extractors with bidirectional recurrent layers and a two-step training scheme on public EEG corpora [25]. **SeqSleepNet** introduced a sequence-to-sequence hierarchy that jointly learns intra-epoch and inter-epoch representations for full-night contextualization [16]. Attention-centric designs such as **AttnSleep** improved feature selectivity and efficiency for single-channel EEG [3]. Transformer pipelines exemplified by **SleepTransformer** brought self-attention, interpretability tools, and uncertainty quantification into end-to-end staging [17]. Architectures such as **CausalAttnNet** have further refined temporal modeling by employing causal convolutions to ensure efficient processing without future information leakage [14]. Notably, this notion of "causality" refers to temporal dependencies in signal processing, which is distinct from the statistical causal inference framework for mitigating confounding that is the focus of our work. In parallel, multi-scale representations with feature pyramids and supervised contrastive learning have enhanced scoring performance; **SleepPyCo** is a recent example reporting improvements across datasets [11].

Despite these advances, supervised pipelines remain sensitive to *dataset shift* arising from differences in sensors/montages, site protocols, labeling procedures, and cohort composition, which can significantly degrade out-of-domain generalization [18]. Establishing true causal relationship between FOI and sleep score so as to enable correct interpretation and robust generalization to heterogeneous environments has therefore become a parallel focus.

2.2 Causal Inference with Continuous Treatments

To build a robust and explainable model, we incorporate causal inference methodologies. The propensity score, namely the probability of treatment assignment, provides a balancing weight for each sample which marginally decorrelates the treatment (which corresponds to FOI in our paper) and potential confounders among observed variables [21]; the propensity score approach has been mainly studied for the binary treatment case and later generalized to generalized propensity score (GPS) for continuous treatments, enabling estimation of dose–response functions (DRFs) [6]. Another line of work avoids modeling the treatment assignment, and directly derives sample weights that enforce zero covariance between the binary treatment and potential confounders while keeping the weights as close to the original distribution as possible, namely *entropy balancing*. The weights can be derived as a solution of

convex program and enjoy favorable asymptotic properties [5, 28]. **Entropy Balancing for Continuous Treatments (EBCT)** [26] adapts this paradigm to continuous treatments [26]. In clustered sleep data—where subject/site/device factors can influence both inputs and labels—such balance-first, model-agnostic weighting is complementary to modern classifiers and facilitates DRF estimation, offering a principled route to robustness and explainability under realistic distribution shifts [23].

Estimating the **dose–response function (DRF)** [20] from observational data is a central challenge in causal inference, often hampered by **confounding** where covariates create spurious correlations between a feature of interest and an outcome [6]. A primary strategy to address this is sample reweighting, with methods evolving from propensity scores [21] to direct balancing approaches. Among these, **Entropy Balancing for Continuous Treatments (EBCT)** [26] is a state-of-the-art technique that finds weights to balance *observed* covariates without needing to model the treatment assignment mechanism [26]. However, a critical limitation of these standard frameworks is their reliance on the **weak unconfoundedness** assumption—that all relevant confounders have been observed. This assumption is often violated in clustered data, where *unobserved* cluster-level factors (e.g., subject-specific physiology) can introduce significant bias. **Our work directly addresses this gap.** We extend the EBCT framework to handle such **unobserved cluster-level confounding** by introducing a novel set of balancing constraints that use an observed cluster indicator (c_i) as a proxy for the unobserved factors (U_i), thereby enabling a more credible and robust estimation of the DRF in complex, realistic settings.

3 Settings

Clustered sleep data and variables

We observe N subjects, each with N_i epochs of 30 s electroencephalogram (EEG). Sleep stages are scored (or mapped) to be compliant with the **AASM** scoring rules, yielding five classes [1]. Thus the sample has a *clustered* structure at the subject level. For epoch j of subject i , we define the following variables, which are summarized in Table 1:

- $\vec{X}_{ij} \in \mathbb{R}^d$: *EEG-derived features*, e.g., Hjorth activity/ mobility/ complexity [7]; absolute/relative band powers and ratios from Welch PSD [27]; spectral entropy.
- $T_{ij} \in \mathbb{R}$: a *continuous FOI*, defined by selecting one component of \vec{X}_{ij} (the remaining indices constitute the \vec{X}_{ij} vector).
- Y_{ij} : observed one-vs-rest indicator for the sleep stage $k \in \{W, N1, N2, N3, REM\}$.

Let c_i denote the observed cluster indicator (subject, and—if available—site/device). We also denote the unobserved, cluster-level factors as U_i (e.g., protocol idiosyncrasies, latent subject traits, characteristics of wearable devices that the subject uses).

Potential outcomes and causal estimates

Let $Y_{ij}(t)$ be the potential outcome [22] for epoch j of subject i when the value of FOI is $t \in \mathcal{T} \subset \mathbb{R}$. Our primary estimate is the

dose–response function (DRF)

$$\mu_k(t) = \frac{1}{N} \sum_{i=1}^N \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbb{E}[Y_{ij}^k(t)],$$

where $Y_{ij}^k(t) = 1\{Y_{ij}(t) = k\}$, $k \in \{W, N1, N2, N3, REM\}$, and thus, $\mathbb{E}[Y_{ij}^k(t)] = \mathbb{P}\{Y_{ij}(t) = k\}$.

Assumptions

We work in the potential-outcomes framework under the following basic assumptions:

- (1) **SUTVA** (Stable Unit Treatment Value Assumption)/**Consistency**: If $T_{ij} = t$, then $Y_{ij} = Y_{ij}(t)$; no interference [22].
- (2) **Ignorability with latent cluster factors**: For all $t \in \mathcal{T}$,

$$Y_{ij}(t) \perp\!\!\!\perp T_{ij} \mid (\vec{X}_{ij}, U_i).$$

This key assumption is visualized in the Directed Acyclic Graph (DAG) in Figure 1. Although U_i is unobserved, our weighting strategy (below) uses the observable proxy c_i to mitigate U_i 's confounding.

- (3) **Positivity (overlap)**: The conditional density of T_{ij} given (\vec{X}_{ij}, c_i) is positive over relevant t .

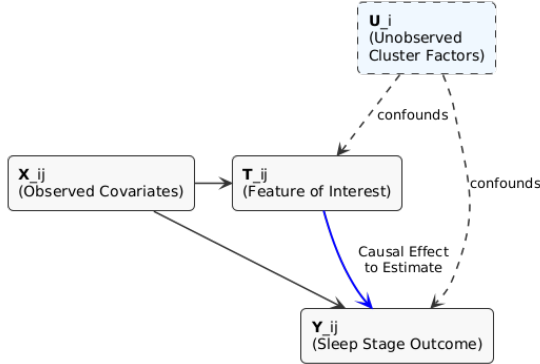


Figure 1: Causal assumptions visualized as a Directed Acyclic Graph (DAG). The diagram shows the **primary causal pathway of interest** from the Feature of Interest to the sleep stage ($T_{ij} \rightarrow Y_{ij}$, blue arrow). It also depicts the confounding "back-door paths" from observed covariates (\vec{X}_{ij}) and unobserved, cluster-level latent factors (U_i), which this framework aims to block.

Construction of balanced pseudo-population

Rather than modeling the treatment mechanism, standard EB/EBCT construct a *balanced pseudo-population* using entropy-based weights $\{\omega_{i,j}\}$ which decorrelate T and \mathbf{X} in the sample so as to balance the distribution of \mathbf{X} according to the value of T , and vice versa [5, 26]. Our extension adds a *within-cluster sign-balance* constraint

$$\sum_{j=1}^{N_i} \omega_{i,j} \text{sign}(T_{ij}^*) \approx 0 \quad (\forall i),$$

where T^* is the standardized T . This constraint reduces treatment–cluster association. Under these balances, DRFs can be estimated on the weighted sample without specifying a propensity model [26, 28]:

$$\hat{\mu}(t) = \frac{1}{\sum_i N_i} \sum_{i,j} \omega_{i,j} \hat{m}(t, \vec{X}_{ij}),$$

where $\hat{m}(t, \vec{X}_{ij})$ is the fitted outcome regression model (e.g., a model that regresses $Y_{ij}^k(T_{ij})$ on T_{ij} and \vec{X}_{ij}).

Signal Processing and Feature Extraction

For each 30-second epoch, we extracted a feature set from the pre-processed EEG signals. The raw signals were first processed according to AASM guidelines, including band-pass filtering (0.3–35 Hz), mains-notching, and artifact removal, before being segmented into 30-second non-overlapping epochs aligned with the hypnogram [1].

From each clean epoch, we derived features from two primary domains. **Spectral features** included absolute and relative powers for standard frequency bands ($\delta, \theta, \alpha, \sigma, \beta$), band ratios, and spectral entropy, computed using Welch's periodogram with a 4-second window and 50% overlap [27]. **Temporal features** were characterized by Hjorth parameters (Activity, Mobility, Complexity) [7]. For the causal analysis, a single index was selected as the continuous FOI (T_{ij}), with the remaining features forming the vector \vec{X}_{ij} . All features were standardized prior to model training as follows,

$$\vec{X}_{ij}^* = \mathbf{S}_X^{-1/2} (\vec{X}_{ij} - \bar{X}), \quad T_{ij}^* = s_T^{-1/2} (T_{ij} - \bar{T}),$$

where $\bar{X} = \frac{1}{N} \sum_{i=1}^N \frac{1}{N_i} \sum_{j=1}^{N_i} \vec{X}_{ij}$ and $\bar{T} = \frac{1}{N} \sum_{i=1}^N \frac{1}{N_i} \sum_{j=1}^{N_i} T_{ij}$,

$$\mathbf{S}_X = \frac{1}{N} \sum_{i=1}^N \frac{1}{N_i} \sum_{j=1}^{N_i} \text{diag}((\vec{X}_{ij} - \bar{X})(\vec{X}_{ij} - \bar{X})^T),$$

$$s_T = \frac{1}{N} \sum_{i=1}^N \frac{1}{N_i} \sum_{j=1}^{N_i} (T_{ij} - \bar{T})^2.$$

Table 1: Summary of Variables and Notations

Symbol	Definition
N	Total number of subjects.
N_i	Number of 30-second epochs for subject i .
i	Subject (cluster) index, $i = 1, \dots, N$.
j	Epoch index for subject i , $j = 1, \dots, N_i$.
\vec{X}_{ij}	EEG-derived features for epoch j of subject i .
T_{ij}	Continuous Feature of Interest (FOI).
Y_{ij}	Observed sleep stage for epoch j of subject i .
k	Sleep stage class $\{W, N1, N2, N3, REM\}$.
c_i	Observed cluster (subject) indicator.
U_i	Unobserved cluster-level latent factors.
$Y_{ij}(t)$	Potential outcome if T_{ij} were set to t .
$\mu_k(t)$	Dose-Response Function (DRF) for stage k .
$\omega_{i,j}$	Entropy-based weight for each epoch.
\vec{X}_{ij}^*, T_{ij}^*	Standardized versions of \vec{X}_{ij}, T_{ij} .

4 Method: EBCT (for i.i.d. samples) and Cluster-Adjusted Extension

Our method is designed to enable causal inference in clustered data settings by extending the Entropy Balancing for Continuous Treatments (EBCT) framework. As illustrated in Figure 2, our approach consists of several key stages, which we detail in the following

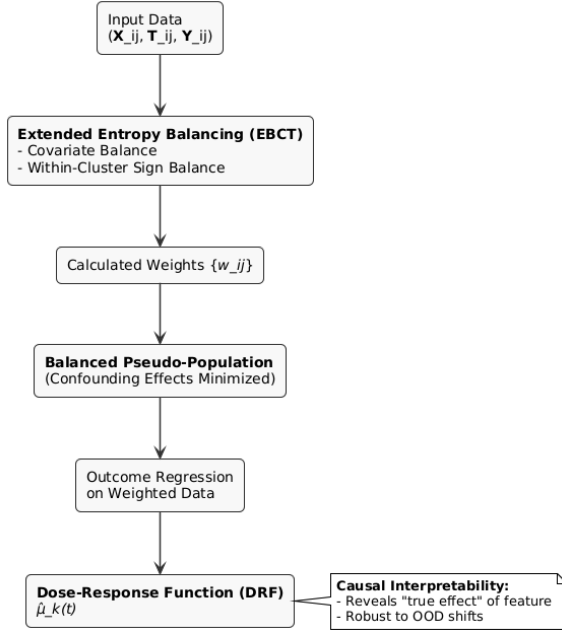


Figure 2: Schematic of the proposed causal framework. The workflow begins with the input data, applies the Extended Entropy Balancing (EBCT) method to generate weights $(\omega_{i,j})$ and a balanced pseudo-population where confounding is minimized. This enables the final estimation of an interpretable Dose-Response Function (DRF).

subsections. We begin by reviewing the original EBCT for i.i.d. samples.

4.1 Original EBCT for i.i.d. samples

The original *Entropy Balancing for Continuous Treatments* (EBCT) of Tübbicke [26] is formulated for i.i.d. samples. We seek nonnegative weights $\omega_{i,j}$ with $\sum_{i=1}^N \sum_{j=1}^{N_i} \omega_{i,j} = 1$ that are as close as possible to the original weights, i.e., the base weights $\{q_{ij} > 0\}$ where $q_{ij} \equiv 1/A$ with $A = \sum_i N_i$.

Since data are not clustered, we flatten the index as $k = 1, \dots, A$ and write (\mathbf{X}_k^*, T_k^*) . EBCT minimizes the Kullback–Leibler divergence from $\{w_k : k = 1, \dots, A\}$ to $\{q_k : k = 1, \dots, A\}$ subject to moment-balance constraints:

$$\begin{aligned}
 & \min_{\{w_k > 0\}} \sum_{k=1}^A w_k \log \left(\frac{w_k}{q_k} \right) \\
 & \text{s.t.} \quad \underbrace{\sum_{k=1}^A w_k T_k^* \mathbf{X}_k^* = \mathbf{0}}_{(1)}, \quad \underbrace{\sum_{k=1}^A w_k \mathbf{X}_k^* = \mathbf{0}}_{(2)}, \quad \underbrace{\sum_{k=1}^A w_k T_k^* = 0}_{(3)}, \\
 & \quad \underbrace{\sum_{k=1}^A w_k = 1}_{(4)}, \quad \underbrace{w_k \geq 0 \quad \forall k}_{(5)}.
 \end{aligned}$$

Here $\mathbf{X}_k^* \in \mathbb{R}^d$, so (1) is a vector constraint (one for each component of \mathbf{X}_k^*).

Covariance decomposition (i.i.d.). Since $\text{Cov}_w(T^*, \mathbf{X}^*) := \mathbb{E}_w[T^* \mathbf{X}^*] - \mathbb{E}_w[T^*] \mathbb{E}_w[\mathbf{X}^*]$ where $\mathbb{E}_w[A] := \frac{1}{\sum_i N_i} \sum_{i=1}^N \sum_{j=1}^{N_i} w_{i,j} A_{i,j}$, constraints (1)–(3) imply $\text{Cov}_w(T^*, \mathbf{X}^*) = \mathbf{0}$, i.e., Pearson decorrelation in the reweighted sample.

KKT and exponential tilting. We solve the convex program using the Lagrange multiplier method. Let γ collect multipliers for (1)–(3) and λ for (4). First-order conditions yield weights of exponential tilting form,

$$w_k = \frac{q_k \exp \left\{ \gamma^\top \begin{bmatrix} T_k^* \mathbf{X}_k^* \\ \mathbf{X}_k^* \\ T_k^* \end{bmatrix} \right\}}{\sum_{\ell=1}^A q_\ell \exp \left\{ \gamma^\top \begin{bmatrix} T_\ell^* \mathbf{X}_\ell^* \\ \mathbf{X}_\ell^* \\ T_\ell^* \end{bmatrix} \right\}},$$

and the value of γ can be obtained by optimizing the strictly concave dual

$$\mathcal{L}_d(\gamma) = -\log \left(\sum_{k=1}^A q_k \exp \left\{ \gamma^\top \begin{bmatrix} T_k^* \mathbf{X}_k^* \\ \mathbf{X}_k^* \\ T_k^* \end{bmatrix} \right\} \right).$$

using the (quasi-)Newton algorithm.¹

4.2 Proposed cluster-adjusted EBCT with within-cluster sign balance

We now propose an extended version of EBCT for clustered settings where clustered level confounders –such as subject characteristics/wearable device characteristics in the sleep scoring example – may not be observed. We *retain* the original EBCT constraints (1)–(5) *unchanged* while adding one per-cluster constraint to mitigate unobserved cluster-level confounding:

$$\begin{aligned}
 & \min_{\{\omega_{i,j} > 0\}} \sum_{i=1}^N \sum_{j=1}^{N_i} \omega_{i,j} \log \left(\frac{\omega_{i,j}}{q_{ij}} \right) \\
 & \text{s.t.} \quad (1) \sum_{i=1}^N \sum_{j=1}^{N_i} \omega_{i,j} T_{ij}^* \vec{\mathbf{X}}_{ij}^* = \mathbf{0}, \\
 & \quad (2) \sum_{i=1}^N \sum_{j=1}^{N_i} \omega_{i,j} \vec{\mathbf{X}}_{ij}^* = \mathbf{0}, \\
 & \quad (3) \sum_{i=1}^N \sum_{j=1}^{N_i} \omega_{i,j} T_{ij}^* = 0, \\
 & \quad (4) \sum_{i=1}^N \sum_{j=1}^{N_i} \omega_{i,j} = 1, \quad (5) \omega_{i,j} \geq 0 \quad \forall i, j, \\
 & \quad (6) \sum_{j=1}^{N_i} \omega_{i,j} \tilde{T}_{ij}^* = 0 \quad (\forall i), \quad (*) \\
 & \quad \tilde{T}_{ij}^* = \mathbf{1}(T_{ij}^* > 0) - \mathbf{1}(T_{ij}^* \leq 0) \in \{-1, +1\}.
 \end{aligned}$$

¹See also entropy balancing in the binary case [5] and its large-sample properties [28].

Interpretation. (1)–(3) enforce $\text{Cov}_w(T_{ij}^*, \vec{X}_{ij}^*) \approx \mathbf{0}$ in the sample. (6) forces the weighted mass of “high” vs. “low” T_{ij}^* to match *within each cluster*, reducing the correlation between T_{ij}^* and the cluster identity. Feasibility of (6) requires within-cluster overlap (both signs present).

KKT with per-cluster multipliers and closed-form profile objective. We again employ the Lagrange multiplier method to solve the convex program. We introduce multipliers γ for (1)–(3), λ for (4), and $\{v_i\}$ for (6). The derived weights retain exponential tilting form

$$\omega_{i,j} \propto q_{ij} \exp \left\{ \gamma^\top \begin{bmatrix} T_{ij}^* \vec{X}_{ij}^* \\ \vec{X}_{ij}^* \\ T_{ij}^* \end{bmatrix} + v_i \tilde{T}_{ij}^* \right\}. \quad (1)$$

We importantly note that for fixed γ , the value of v_i has closed form. Define

$$A_i = \sum_{j: \tilde{T}_{ij}^* = +1} q_{ij} \exp \left\{ \gamma^\top \begin{bmatrix} T_{ij}^* \vec{X}_{ij}^* \\ \vec{X}_{ij}^* \\ T_{ij}^* \end{bmatrix} \right\}, \quad B_i = \sum_{j: \tilde{T}_{ij}^* = -1} q_{ij} \exp \left\{ \gamma^\top \begin{bmatrix} T_{ij}^* \vec{X}_{ij}^* \\ \vec{X}_{ij}^* \\ T_{ij}^* \end{bmatrix} \right\}.$$

Then

$$v_i(\gamma) = \frac{1}{2} \log \frac{B_i}{A_i}$$

Plugging in $v_i(\gamma)$ into (1) gives a profile dual objective with respect to γ ,

$$\mathcal{L}_d(\gamma) = -\log Z(\gamma) = -\log 2 - \log \left(\sum_{i=1}^N \sqrt{A_i B_i} \right),$$

where

$$Z(\gamma, \{v_i\}) = \sum_{i=1}^N \sum_{j=1}^{N_i} q_{ij} \exp \left\{ \gamma^\top \begin{bmatrix} T_{ij}^* \vec{X}_{ij}^* \\ \vec{X}_{ij}^* \\ T_{ij}^* \end{bmatrix} + v_i \tilde{T}_{ij}^* \right\} = \sum_{i=1}^N 2 \sqrt{A_i(\gamma) B_i(\gamma)}.$$

This profile objective remains strictly concave in γ and hence can be optimized using relevant optimization algorithm.

4.3 Diagnostics and downstream use

Balance checks. After solving the convex program, we verify whether the following constraints are satisfied in the sample before using the weights in downstream tasks.

$$\begin{aligned} \sum_{i,j} \omega_{i,j} T_{ij}^* \vec{X}_{ij}^* &\approx \mathbf{0}, & \sum_{i,j} \omega_{i,j} \vec{X}_{ij}^* &\approx \mathbf{0}, \\ \sum_{i,j} \omega_{i,j} T_{ij}^* &\approx 0, & \sum_{j=1}^{N_i} \omega_{i,j} \tilde{T}_{ij}^* &\approx 0 \quad (\forall i). \end{aligned}$$

Downstream tasks. We specify two downstream tasks. We use $\{\omega_{i,j}\}$ to (i) estimate the DRF $\mu(t)$ by weighted regression of Y_{ij} on \vec{X}_{ij} and T_{ij} and (ii) fit classifiers (e.g., logistic/RF) for sleep staging on the reweighted sample. We explain each task in detail in Sections 4.4 and 4.5 respectively.

4.4 Dose–Response Function (DRF): Definition and Estimation

Definition. Let $Y_{ij}(t)$ denote the potential outcome of epoch subject i at epoch j when the value of FOI or treatment is equal to $t \in \mathcal{T} \subset \mathbb{R}$. The *dose–response function (DRF)* [20] is

$$\mu_k(t) = \frac{1}{N} \sum_{i=1}^N \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbb{E}[Y_{ij}^k(t)],$$

which maps each dose level t to the expected outcome if, counterfactually, all units were assigned $T=t$.

Balance-based identification. Under SUTVA, overlap, and weak unconfoundedness $Y_{ij}^k(t) \perp\!\!\!\perp T_{ij} \mid (\vec{X}_{ij}, c_i)$, the EBCT constraints (1)–(6) create a *balanced pseudo-population* that attenuates FOI–X and FOI–cluster associations. Outcome models fitted on this reweighted sample can recover $\mu_k(t)$ over regions of adequate overlap, without specifying a parametric treatment model [5, 26, 28].

Feature map for DRF models. We use the following basis to capture nonlinear dose effects and dose–confounder interactions:

$$f(t, x) = (1, t^*, t^{*2}, x^{*\top}, (t^* x^*)^\top, (t^{*2} x^*)^\top)^\top,$$

where $t^* = s_T^{-1/2}(t - \bar{T})$ and $x^* = \vec{X}_{ij}^*$. When reporting $\mu(t)$ on the original scale, we transform $t \mapsto t^*$ before evaluation.

DRF estimation. For each class $k \in \{W, N1, N2, N3, \text{REM}\}$ we estimate the class probability curve $\mu_k(t)$ using logistic regression. Denote the logistic function as $\sigma(u) = (1 + e^{-u})^{-1}$. We estimate the regression parameter β_k by weighted maximum likelihood:

$$\begin{aligned} \hat{\beta}_k &= \arg \min_{\beta_k} \sum_{i=1}^N \sum_{j=1}^{N_i} \omega_{i,j} \left\{ -Y_{ij}^k \log \pi_{ij}^k - (1 - Y_{ij}^k) \log (1 - \pi_{ij}^k) \right\}, \\ \pi_{ij}^k &= \sigma(f(T_{ij}^*, \vec{X}_{ij}^*)^\top \beta_k). \end{aligned}$$

The DRF for class k is then computed by averaging the estimated probabilities over all data points for each dose value t :

$$\hat{\mu}_k^{\text{GLM}}(t) = \sum_{i=1}^N \sum_{j=1}^{N_i} \omega_{i,j} \sigma(f(t^*, \vec{X}_{ij}^*)^\top \hat{\beta}_k).$$

Vector-valued DRF and normalization. Repeating the DRF estimation for each class k , we obtain $\hat{\mu}(t) = (\hat{\mu}_W(t), \hat{\mu}_{N1}(t), \dots, \hat{\mu}_{\text{REM}}(t))$. With one-vs-rest fitting as described above, $\sum_k \hat{\mu}_k(t)$ is not necessarily exactly equal to 1; if desired, we can report normalized probabilities $\tilde{\mu}_k(t) = \hat{\mu}_k(t) / \sum_{k'} \hat{\mu}_{k'}(t)$ over the region where the denominator is nonzero.

Uncertainty quantification via cluster-level bootstrap. To quantify the uncertainty of our DRF estimates, we employ a cluster bootstrap procedure. This involves repeating the entire estimation process on bootstrapped samples: (1) resample clusters i with replacement to create a new dataset, (2) recompute the weights $\{\omega_{i,j}^{(b)}\}$ for this new dataset, and (3) refit the model to form the DRF estimate $\hat{\mu}_k^{(b)}(t)$. From the collection of B bootstrap estimates, we report the mean, standard error, and pointwise 95% confidence intervals:

$$\bar{\mu}_k(t) = \frac{1}{B} \sum_{b=1}^B \hat{\mu}_k^{(b)}(t), \quad \widehat{\text{SE}}_k(t) = \sqrt{\frac{1}{B-1} \sum_{b=1}^B (\hat{\mu}_k^{(b)}(t) - \bar{\mu}_k(t))^2},$$

and pointwise 95% CIs as $\bar{\mu}_k(t) \pm 1.96 \widehat{SE}_k(t)$.

4.5 Weighted Classification for Robust Sleep Staging

In parallel with the causal interpretation via DRFs, we train a robust multi-class sleep-stage classifier on the *reweighted* sample. Let T_{ij}^* denote the standardized FOI and $\vec{X}_{ij}^* \in \mathbb{R}^d$ the standardized potential confounders (Sec. 3). We use a Random Forest (RF) as the base classifier.

Weighted training. We fit the RF with the **cluster-adjusted EBCT** sample weights $\omega_{i,j}$. Thus, tree splits and node impurities are computed with respect to the balanced pseudo-population induced by $\omega_{i,j}$. We denote the RF’s class-probability output by $\hat{\pi}_k^{\text{RF}}(T_{ij}^*, \vec{X}_{ij}^*)$, $k \in \{W, N1, N2, N3, \text{REM}\}$.

Given a new input (t^*, x^*) , the trained forest returns class probabilities by tree averaging, where M is the number of trees:

$$\hat{\pi}_k^{\text{RF}}(t^*, x^*) = \frac{1}{M} \sum_{m=1}^M \hat{\pi}_k^m(t^*, x^*),$$

and predicts the class which has the largest probability as follows,

$$\hat{Y}_{ij}(t^*, x^*) = \underset{k}{\text{argmax}} \hat{\pi}_k^{\text{RF}}(t^*, x^*).$$

Evaluation. We report **Accuracy**, **Macro-F1** and **Cohen’s Kappa** (κ) and additionally, per-class F1 for N1/REM. Comparisons are made with the RF trained on original samples without applying any weights, which cannot benefit from the effect of balancing.

5 Experiments

5.1 Datasets and Cluster Structure

We evaluate our proposed framework on the Sleep-EDF database *Expanded* (version 2018). This database contains 197 whole-night PolySomnoGraphy (PSG) sleep recordings, containing EEG, EOG, chin, EMG, and event markers. For this study, we use the **Sleep Cassette (SC)** subset (commonly referred to as *Sleep-EDFx-78*). The SC subset comprises overnight PSG from **78** healthy subjects (age **25–101** years), totaling **153** nights/recordings [9, 12]. The sleep-edf database contains Hypnograms were manually scored according to the rules [19]. While, the PSG data contains multiple signals, our proposed framework utilizes only the EEG signals for sleep stage classification. Following previous works [11], the duration of an EEG epoch was set to 30 s. Furthermore, all EEG signals were sampled at **100 Hz**.

Train-Test Split. To evaluate the model’s generalization performance on unseen subjects, we partitioned the dataset at the subject level. Specifically, we randomly selected **20 subjects for the training set**, which was used to estimate the EBCT weights and train the classifiers. The remaining **5 subjects were held out as a test set** for final evaluation. Across subjects, the number of 30 s epochs per selected 25 subject ranged from **1,027** to **4,984** (mean \approx **2,467**). This strict separation ensures that no data from the test subjects is used during any part of the training process.

Label mapping and preprocessing. Following common practice for Sleep-EDF, we map R&K labels $\{W, 1, 2, 3, 4, R, M, ?\}$ into the

Table 2: Sleep-EDF-2018 class distribution after preprocessing (five-class mapping, night-only, W 30 min trimming, M/? removed).

Stage	Epochs	Share
Wake (W)	69,824	35.0%
N1	21,522	10.8%
N2	69,132	34.68%
N3 (S3US4)	13,039	6.54%
REM	25,835	12.96%
Total	199,352	100%

five-class scheme $\{W, N1, N2, N3, \text{REM}\}$ by *merging classes 3 and 4 into N3* and discarding M and $?$ [4]. To mitigate long Wake segments before/after the night and reduce imbalance, we apply *W-trimming*: retain only the first/last **30 min** of Wake around the sleep period [4].

Table 2 reports the resulting class distribution on Sleep-EDF-2018 (five-class mapping, night-only, 30 min W-trimming, $M/?$ removed).

5.2 Signal Processing and Features

All recordings are filtered (e.g., 0.3–35 Hz band-pass and mains-notch), re-referenced, and segmented into 30 s epochs aligned with the hypnogram. For each epoch (i, j) , we compute a feature vector \vec{X}_{ij} comprising standard EEG-derived indices (e.g., Hjorth activity/mobility/complexity, absolute/relative band powers $\delta, \theta, \alpha, \sigma, \beta$, band ratios, spectral entropy). The *continuous FOI* T_{ij} is defined by *selecting one* index from \vec{X}_{ij} (e.g., Hjorth mobility or relative δ power). We standardize to $(\vec{X}_{ij}^*, T_{ij}^*)$ as per Section 3.

5.3 Baselines

We compare our proposed method against a No-weight baseline, which represents a standard model trained on the original, unweighted sample. In contrast, our proposed method applies the EBCT framework by incorporating an additional within-cluster sign-balance constraint (constraint $(*)$) to explicitly account for both the observed confounders and unobserved cluster-level confounders.

5.4 Implementation Details

Causal weight estimation (CVXPY/SCS). We estimate EBCT weights $\{\omega_{i,j}\}$ by solving the convex program proposed in Section 4.2. Optimization uses the **SCS** conic solver [13] with tolerance 10^{-6} ; we apply warm starts across CV folds and bootstrap replicates.

Random Forest Classifier. We implement the Random Forest classifier using the `scikit-learn` library [15]. We adopt the default hyperparameter settings, except for fixing the number of trees ($M=200$). All other parameters (split criteria, depth, and regularization terms) follow the library defaults.

5.5 Results

5.5.1 Classification Performance. Table 2 reports **Macro-F1** as primary metrics, with **Accuracy** as secondary. We also show **Cohen’s**

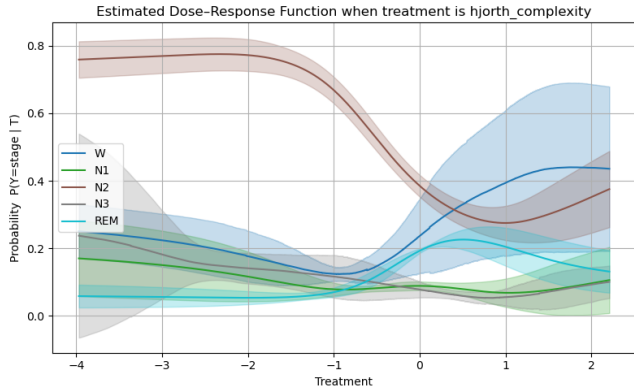


Figure 3: GLM-based class-wise DRF for *Hjorth complexity* (EBCT-weighted fit). Shaded bands denote 95% subject-bootstrap CIs ($B = 20$).

Kappa (κ), per-class F1 for N1 and REM. Table 4 reports per-class recall values as well.

Table 3: Random Forest overall performance before (No-Weight) and after applying our cluster-adjusted weights (Proposed).

Method	Acc	Macro-F1	κ	F1(N1)	F1(REM)
No-Weight	0.681	0.522	0.546	0.099	0.562
Proposed	0.691	0.540	0.561	0.086	0.568

Table 4: Per-class Recall comparison for Random Forest, highlighting the improvement in deep sleep (N3) detection.

Method	Recall(W)	Recall(N1)	Recall(N2)	Recall(N3)	Recall(REM)
Unweighted	0.957	0.060	0.733	0.414	0.580
Proposed	0.963	0.050	0.743	0.569	0.577

5.5.2 Balance Diagnostics (numeric summary). To verify the effectiveness of our weighting method, we compare the covariance between X and T in the original samples and the reweighted samples obtained using cluster-adjusted EBCT weights. Specifically, for each FOI, we report the maximum covariance (taken over the d variables in X) before and after EBCT weighting, as summarized in Table 5.

$$\max|\text{cov}| = \max_{1 \leq \ell \leq d} \text{cov}_w(T_{ij}^*, \vec{X}_{ij\ell}^*).$$

5.5.3 Dose-Response Functions (DRFs). Figure 3 shows the DRF function when the FOI is Hjorth complexity. Interpretation is restricted to regions with adequate overlap. Shaded areas in the figures denote 95% CIs computed from the subject-level bootstrap ($B = 20$). The x-axis is the treatment value (FOI intensity); the y-axis is $\mathbb{P}\{Y_{ij}(t) = k\}$.

Table 5: Balance before vs. after EBCT (lower is better).

FOI	Unweighted	EBCT-weighted
	$\max \text{cov} $	$\max \text{cov}_w $
Hjorth complexity	1.015	4.50e-08
Hjorth activity	1.528	4.04e-06
Alpha ratio	4.183	1.51e-06
Spindle power	1.595	3.67e-07
Spectral entropy	4.938	3.66e-05
Spindle ratio	3.273	7.87e-06
Overall (avg)	2.755	8.41e-06

6 Discussion

Summary of Key Findings. Our main findings are summarized in Table 3 and Table 4. As shown in Table 3, our proposed Random Forest model using cluster-adjusted EBCT weights ('Proposed') achieved a modest but consistent improvement in overall performance metrics compared to the No-Weight baseline, including Accuracy (0.691 vs. 0.681), Macro-F1 (0.540 vs. 0.522), and Cohen's Kappa (0.561 vs. 0.546).

The Value of Causal Interpretability. A key contribution of this work is its potential to interpret the role of specific EEG features, moving beyond simple prediction. While the roles of specific frequency bands like delta or alpha are well-established, the influence of composite indices such as *Hjorth complexity*—which reflects signal irregularity—on sleep stages is less clear. Our causal inference framework, by controlling for both observed confounders and unobserved cluster-level confounder, provides a tool to explore how the intensity of such complex features might affect the probability of each sleep stage via 'what-if' scenarios (DRF: Fig 3). Unlike conventional models that rely solely on correlation, this opens up an avenue for generating new physiological hypotheses directly from data.

Robustness Over Raw Performance. It is noteworthy that the predictive performance was maintained or slightly improved after applying causal weights. This has a more significant implication than a mere increase in accuracy. A model with a reduced dependency on spurious correlations with confounders is less likely to overfit to the idiosyncrasies of the training dataset. Therefore, it can be expected to exhibit more stable and robust performance on new data from different measurement environments or subject groups.

Limitations and Future Work. This study has several limitations. First, the number of subjects used for training was 20 out of a total of 78 (Sec 5.1), so caution is needed when generalizing the relationships found in this study. Second, the weight estimation process using EBCT with CVXPY is computationally expensive, making it challenging to apply directly to larger datasets. Future work should aim to validate these findings on a larger scale and explore more efficient optimization algorithms.

In addition to practical limitations, there are also methodological considerations. A key methodological aspect of our work is the use of observed cluster indicator, c_i , as a proxy for unobserved

cluster-level factors, U_i . While this approach pragmatically addresses confounding arising from unobserved cluster-level factors, as supported by improved OOD performance in our experiments, a formal justification of sufficiency remains an open question. Future work could involve sensitivity analyses to assess the proxy's validity and theoretical investigations into the conditions under which our constraint guarantees unbiasedness.

7 Conclusion

This study presented a classification model for sleep staging that, through EBCT-based causal weighting, is both robust to confounders (both observed and unobserved cluster-level ones) and interpretable, all while maintaining predictive performance. Notably, the framework demonstrated the potential for causal exploration even for EEG features like Hjorth_complexity whose roles are not yet clearly defined. Although the limitations of sample size and computational cost are apparent, this research offers an important direction for sleep analysis—moving beyond simple prediction toward robust and reliable physiological insights.

Acknowledgments

This research was supported by Korean Institute for Advancement of Technology (KIAT) grant funded by the Korea Government (MOTIE) (RS-2024-00435815, Human Resource Development Program for Industrial Innovation (Global)).

References

- [1] Richard B Berry. 2014. The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications. version 2.1. *Darien Illinois: American Academy of Sleep Medicine* (2014).
- [2] Massimiliano de Zambotti, Nicola Cellini, Aimée Goldstone, Ian M. Colrain, and Fiona C. Baker. 2019. Wearable Sleep Technology in Clinical and Research Settings. *Medicine & Science in Sports & Exercise* 51, 7 (2019), 1538–1557. doi:10.1249/MSS.0000000000001947
- [3] Emadeldeen Eldele, Zhenghua Chen, Chengyu Liu, Min Wu, Chee-Keong Kwoh, Xiaoli Li, and Cuntai Guan. 2021. An attention-based deep learning approach for sleep stage classification with single-channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 29 (2021), 809–818.
- [4] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti S. Hämäläinen. 2013. MEG and EEG Data Analysis with MNE-Python. *Frontiers in Neuroscience* 7, 267 (2013), 1–13. doi:10.3389/fnins.2013.00267
- [5] Jens Hainmueller. 2012. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis* 20, 1 (2012), 25–46.
- [6] Keisuke Hirano and Guido W. Imbens. 2004. *The Propensity Score with Continuous Treatments*. John Wiley Sons, Ltd, Chapter 7, 73–84. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/0470090456.ch7 doi:10.1002/0470090456.ch7
- [7] Bo Hjorth. 1970. EEG analysis based on time domain properties. *Electroencephalography and clinical neurophysiology* 29, 3 (1970), 306–310.
- [8] Institute of Medicine (US) Committee on Sleep Medicine and Research. 2006. *Sleep Disorders and Sleep Deprivation: An Unmet Public Health Problem*. National Academies Press, Washington, DC. doi:10.17226/11617
- [9] Bob Kemp, Aeilko H Zwinderman, Bert Tuk, Hilbert AC Kamphuisen, and Josefin JL Obery. 2000. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering* 47, 9 (2000), 1185–1194.
- [10] S. Khosla, M. C. Deak, D. Gault, and et al. 2018. Consumer Sleep Technology: An American Academy of Sleep Medicine Position Statement. *Journal of Clinical Sleep Medicine* 14, 5 (2018), 877–880. doi:10.5664/jcsm.7128
- [11] Seongju Lee, Yeonguk Yu, Seunghyeok Back, Hogeon Seo, and Kyoobin Lee. 2024. SleepPyCo: Automatic sleep scoring with feature pyramid and contrastive learning. *Expert Systems with Applications* 240 (2024), 122551. doi:10.1016/j.eswa.2023.122551
- [12] Chengfan Li, Yueyu Qi, Xuehai Ding, Junjuan Zhao, Tian Sang, and Matthew Lee. 2022. A deep learning method approach for sleep stage classification with EEG spectrogram. *International journal of environmental research and public health* 19, 10 (2022), 6322.
- [13] Brendan O'donoghue, Eric Chu, Neal Parikh, and Stephen Boyd. 2016. Conic optimization via operator splitting and homogeneous self-dual embedding. *Journal of Optimization Theory and Applications* 169, 3 (2016), 1042–1068.
- [14] Jie Pan, Yongjie Feng, Pengjun Zhao, Xiaoyu Zou, Aiping Hou, and Xiaoyi Che. 2024. Causalattennet: A fast and long-term-temporal network for automatic sleep staging with single-channel eeg. *IEEE Transactions on Instrumentation and Measurement* (2024).
- [15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
- [16] Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y Chén, and Maarten De Vos. 2019. SeqSleepNet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 27, 3 (2019), 400–410.
- [17] Huy Phan, Kaare Mikkelsen, Oliver Y Chén, Philipp Koch, Alfred Mertins, and Maarten De Vos. 2022. Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification. *IEEE Transactions on Biomedical Engineering* 69, 8 (2022), 2456–2467.
- [18] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence (Eds.). 2009. *Dataset Shift in Machine Learning*. MIT Press, Cambridge, MA.
- [19] Allan Rechtschaffen. 1968. A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. (1968), 1–55.
- [20] James M Robins, Miguel Angel Hernan, and Babette Brumback. 2000. Marginal structural models and causal inference in epidemiology. 550–560 pages.
- [21] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55.
- [22] Donald B Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* 66, 5 (1974), 688.
- [23] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Toward causal representation learning. *Proc. IEEE* 109, 5 (2021), 612–634.
- [24] Sharon Schutte-Rodin, Maryann C Deak, Seema Khosla, Cathy A Goldstein, Michael Yurcheshen, Ambrose Chiang, Dominic Gault, Joseph Kern, Daniel O'Hearn, Scott Ryals, et al. 2021. Evaluating consumer and clinical sleep technologies: an American Academy of Sleep Medicine update. *Journal of Clinical Sleep Medicine* 17, 11 (2021), 2275–2282.
- [25] Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. 2017. DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG. *IEEE transactions on neural systems and rehabilitation engineering* 25, 11 (2017), 1998–2008.
- [26] Stefan Tübbicke. 2022. Entropy balancing for continuous treatments. *Journal of Econometric Methods* 11, 1 (2022), 71–89.
- [27] Peter Welch. 2003. The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics* 15, 2 (2003), 70–73.
- [28] Qingyuan Zhao and Daniel Percival. 2017. Entropy balancing is doubly robust. 20160010 pages.