CROSS-MODAL PRETRAINING WITH MODALITY-SPECIFIC RISK PREDICTION FROM POLYSOMNOGRAPHY

Anonymous authors
Paper under double-blind review

ABSTRACT

Polysomnography (PSG) records multiple physiological systems overnight and may contain prognostic information beyond sleep diagnosis. Using these signals for future disease prediction is challenging because PSG modalities differ in signal structure, availability, and relevance to individual outcomes. Although cross-modal learning can exploit shared physiology, merging all modalities into a single downstream representation may obscure information specific to individual physiological systems.

We introduce a multimodal self-supervised framework that separates cross-modal pretraining from downstream risk integration. Independently parameterized encoders are trained using within-modality masked latent prediction and cross-modal prediction from other simultaneously observed signals. For time-to-event prediction, the encoders are frozen, and endpoint-specific whole-night aggregators and risk heads estimate risk separately from each available modality. These estimates are combined only at the prediction stage, allowing inference from incomplete modality sets without imputing missing signals or representations.

We evaluate ten incident disease outcomes in 26,782 participants from three development cohorts and six harmonized outcomes in an external cohort of 1,000 participants. In pooled evaluation, combining PSG with basic clinical covariates achieves a mean Harrell's C-index of 0.717. Although selected baseline performs better under the matched frozen probe, endpoint-specific downstream modeling increases the mean C-index of our PSG representation from 0.681 to 0.696 and yields higher point estimates than the SleepFM probe for eight of ten outcomes. Representation similarity exceeds the permutation null for 14 of 15 modality pairs, yet the predictive contribution of each modality varies across endpoints. Leave-one-cohort-out, external-cohort, and modality-removal evaluations further assess transfer under heterogeneous input configurations. These findings show that multimodal PSG can benefit from cross-modal learning without requiring physiological modalities to share a single downstream representation.

1 INTRODUCTION

Polysomnography (PSG) is the gold-standard tool for clinical sleep assessment, recording neural, cardiac, respiratory, oxygenation, and muscle activity overnight (Kapur et al., 2017).Beyond sleep diagnosis, PSG-derived measures of hypoxic burden, heart-rate variability, and EEG brain age have been associated with cardiovascular events and long-term health outcomes(Azarbarzin et al., 2019; 2020; Paixao et al., 2020; Zhang et al., 2020). Routinely collected PSG may therefore support prospective disease-risk prediction. The central challenge is that PSG combines signals from distinct physiological systems, while the available channels vary across cohorts and recording settings.

Multimodal learning can capture shared information across signals, but forcing their representations to be similar may discard modality-specific features useful for predicting disease risk.(Federici et al., 2020; Tian et al., 2020b). This issue is especially relevant to PSG, its modalities are synchronized and physiologically related, yet they measure different processes whose predictive value may vary by disease. SleepFM demonstrates the value of cross-modal pretraining and large-scale PSG repre-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

sentation learning (Thapa et al., 2024; 2026), but how such representations should be combined for multi-disease time-to-event prediction remains unclear.

We propose a multimodal self-supervised framework that separates cross-modal representation learning from downstream risk integration. Each physiological modality has an independently parameterized encoder. During pretraining, masked latent prediction models structure within a modality, while cross-modal prediction estimates a target representation from the other observed modalities. For each disease endpoint, frozen representations are summarized by modality-specific whole-night aggregators and converted into separate risk scores. The scores of the observed modalities are combined only at the final prediction stage; unavailable modalities are omitted rather than imputed.

We study this question with a multimodal self-supervised framework that separates cross-modal representation learning from downstream risk integration. Each physiological modality is modeled by an independently parameterized encoder with an input configuration suited to its signal characteristics.

We evaluate the framework using three development cohorts and a held-out subset of an external cohort. The proposed full model achieves a pooled macro-averaged Harrell's C-index of **0.717** (95% CI, 0.702–0.733) for multi-endpoints prediction, including 0.749 for atrial fibrillation, 0.751 for acute kidney injury, and 0.747 for chronic kidney disease. Its average improvement over the basic clinical baseline remains positive in both leave-one-cohort-out evaluations. Representation similarity exceeds the permutation null for 14 of 15 modality pairs, while fusion and modality-removal analyses reveal endpoint-dependent contributions from individual physiological systems. These results support our central design: cross-modal pretraining captures shared physiological structure, whereas modality-specific risk pathways preserve endpoint-relevant information and allow prediction from the modalities available at inference. By extracting prognostic information beyond age and sex from routinely collected PSG, the framework may support disease risk stratification without requiring additional physiological measurements.

The proposed work makes the following contributions:

1. We introduce a self-supervised framework that learns within- and cross-modal PSG structure while retaining modality-specific pathways for time-to-event prediction.
2. Through controlled pretraining and downstream ablations, representation-similarity analysis, and modality-removal experiments, we show that PSG modalities share patient-linked structure but are not interchangeable for disease-risk prediction.
3. We evaluate the framework across three cohorts, including leave-one-cohort-out, external-cohort, and missing-modality settings, and assess whether PSG-derived risk adds information beyond basic clinical context.

2 Related Work

2.1 Learning from physiological and sleep signals

Self-supervised learning has been used to learn physiological representations without dense labels. CLOCS contrasts ECG signals across leads, time segments, and patients, improving label-efficient downstream classification (Kiyasseh et al., 2021). SleepFM extends this direction to multimodal PSG and future disease-risk prediction under heterogeneous channel configurations (Thapa et al., 2024; 2026). We instead focus on how within- and cross-modal pretraining should be coupled with downstream survival modeling.

2.2 Shared and modality-specific information

Multiview methods often learn information shared across views (Tian et al., 2020a), while other multimodal models retain fine-grained unimodal representations alongside shared embeddings (**?**). Information unique to one view can be discarded only when the views are sufficiently redundant for the target task; invariance should otherwise preserve task-relevant information (**??**). Motivated by this distinction, we compare shared feature representations with modality-specific risk pathways for PSG.

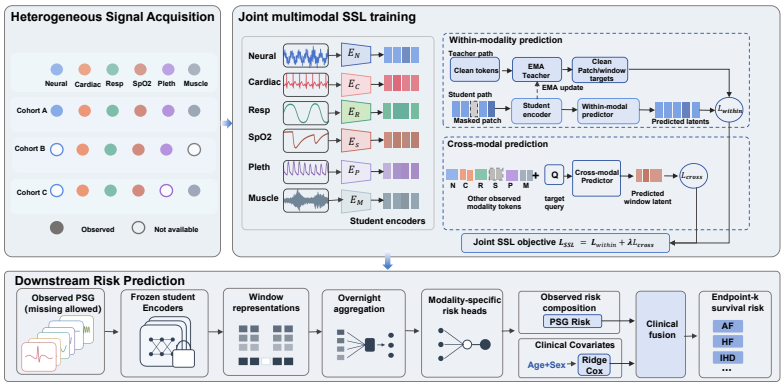

Figure 1: Model overview. Modality-specific encoders are pretrained using within-modality masked prediction and cross-modal latent prediction against EMA-teacher targets. The frozen encoders feed endpoint-specific aggregators and risk heads. Scores from observed modalities are combined without imputation and fused with an age-and-sex score using a validation-selected weight.

2.3 Prediction with missing modalities

Incomplete multimodal inputs have been addressed through methods such as product-of-experts inference and training over subsets of observed modalities (Wu & Goodman, 2018). SleepFM also supports heterogeneous PSG channel configurations (Thapa et al., 2026). Our downstream model instead composes risk estimates from the modalities that are present, without imputing signals or latent features, and we evaluate fixed models under prespecified test-time modality removal.

3 Method

3.1 PSG Inputs and Modality-Specific Encoders

We group PSG channels into six modalities: neural (EEG/EOG), cardiac (ECG), ventilation (airflow, respiratory effort, and snoring), oxygen saturation (SpO_2), plethysmography, and muscle activity (EMG). Recordings are divided into 30-s windows, and channel- and window-level masks identify valid inputs. Full channel mappings, sampling rates, quality control, and cohort-specific preprocessing are described in Section 4 and the appendix.

Each modality m is processed by an independently parameterized encoder E_m. The encoders share an architecture family but use signal-specific input configurations and do not share weights. Each encoder maps an observed 30-s waveform window to patch-level features and a contextualized window representation; channels of the same signal type may share an input stem within an encoder. Cross-modal information is introduced by the pretraining objective rather than by parameter sharing.

3.2 Joint Self-Supervised Pretraining

For each modality, a student encoder processes perturbed inputs, while an exponential-moving-average (EMA) teacher encodes the corresponding clean inputs. Selected patch embeddings are replaced with learned mask tokens before student contextualization. We use sg for stop-gradient, LN for layer normalization, and ℓ for Smooth-L1 loss.

Within-modality prediction. Following the latent-prediction principle of data2vec (Baevski et al., 2022), separate predictors estimate clean teacher representations at masked channel-patch and modality-context positions. A window-level consistency term additionally aligns the student and

teacher window representations. For modality m, the objective is

$$\begin{aligned}\mathcal{L}_{\text{within}}^{(m)} = \ell_{\text{mask}}\left(\hat{z}_m^{\text{ch}}, \text{sg}\left[\text{LN}\left(\bar{z}_m^{\text{ch}}\right)\right]\right) \\ + \ell_{\text{mask}}\left(\hat{z}_m^{\text{ctx}}, \text{sg}\left[\text{LN}\left(\bar{z}_m^{\text{ctx}}\right)\right]\right) \\ + \lambda_v \ \text{MSE}\left(\text{LN}(u_m), \text{sg}[\text{LN}(\bar{u}_m)]\right).\end{aligned} \tag{1}$$

Here, $\hat{z}$ is a student prediction, $\bar{z}$ is its clean teacher target, and u_m and $\bar{u}_m$ are the student and teacher window representations. The masked losses are averaged over eligible masked positions, and the consistency term is averaged over valid windows.

Cross-modal prediction. Unlike missing-signal reconstruction, cross-modal prediction does not create synthetic inputs or fill absent modalities. It only uses truly observed signals during pretraining. For an observed target modality m, a learned query attends to student window representations from the other observed modalities in the same 30-s window. The target modality is excluded from its own context. Context dropout may remove additional modalities during training while retaining at least one. Let $\mathcal{C}_{i,t,m}$ be the retained context set. The target prediction is

$$\hat{u}_{i,t,m} = P_m\left(q_m, \{u_{i,t,j} : j \in \mathcal{C}_{i,t,m}\}\right), \qquad \mathcal{C}_{i,t,m} \subseteq \mathcal{O}_{i,t} \setminus \{m\}. \tag{2}$$

The predictor P_m consists of a shared cross-modal attention module and a target-specific output head.

For modality m, let $\mathcal{V}_m$ contain windows with an observed target and at least one retained context modality. The cross-modal loss is

$$\mathcal{L}_{\text{cross}} = \underset{m}{\text{mean}} \underset{(i,t)\in\mathcal{V}_m}{\text{mean}} \ \ell\left(\text{LN}(\hat{u}_{i,t,m}), \text{sg}[\text{LN}(\bar{u}_{i,t,m})]\right). \tag{3}$$

The inner mean is computed over valid windows for each target modality, followed by an equal-weight mean across eligible target modalities. Cross-modal predictions provide supervision only during pretraining and are not used to impute missing downstream inputs.

The complete pretraining objective is

$$\mathcal{L}_{\text{SSL}} = \underset{m}{\text{mean}} \, \mathcal{L}_{\text{within}}^{(m)} + \lambda_c(r)\mathcal{L}_{\text{cross}}, \tag{4}$$

where r denotes the optimization step and $\lambda_c(r)$ is warmed up during training. Student encoders and prediction heads are optimized by backpropagation, while teacher parameters are updated by EMA. After pretraining, only the student encoders are retained for downstream prediction.

3.3 Modality-Specific Survival Prediction

The pretrained student encoders are frozen during downstream training. Clean PSG windows are encoded, pooled into fixed 5-min bins, and standardized using downstream training-set statistics. Let $F_{i,m}$ denote the resulting sequence of valid features for participant i and modality m.

For endpoint k, a modality-specific aggregator $A_{m,k}$ applies an input projection, attention-weighted mean-and-standard-deviation pooling, and an output projection. A linear head $R_{m,k}$ then produces a scalar risk score:

$$h_{i,m,k} = A_{m,k}(F_{i,m}), \qquad \eta_{i,m,k} = R_{m,k}(h_{i,m,k}). \tag{5}$$

The final PSG score is the mean of the available modality scores:

$$\eta_{i,\text{PSG},k} = \frac{1}{|\mathcal{O}_i|} \sum_{m\in\mathcal{O}_i} \eta_{i,m,k}. \tag{6}$$

Here, $\mathcal{O}_i$ contains the modalities with usable downstream features for participant i. Missing modalities are excluded from the sum and its denominator.

For each endpoint, the aggregators and risk heads are trained jointly using

$$\begin{aligned}\mathcal{L}_{\text{down},k} = \mathcal{L}_{\text{Cox}}\left(\eta_{\text{PSG},k}; \mathcal{D}_k\right) \\ + \lambda_a \underset{m\in\mathcal{M}_k}{\text{mean}} \, \mathcal{L}_{\text{Cox}}\left(\eta_{m,k}; \mathcal{D}_{m,k}\right).\end{aligned} \tag{7}$$

Here, $\mathcal{D}_k$ is the endpoint-specific training set, $\mathcal{D}_{m,k}$ is its subset with modality m available, and $\mathcal{M}_k$ contains the modalities with available samples and at least one observed event. We use the Breslow approximation for tied event times and fit a separate downstream model for each endpoint.

3.4 Clinical Integration

To assess whether PSG adds predictive information beyond age and sex, we fit a separate ridge-penalized Cox model using these two covariates. Its score is combined with the PSG score after separate standardization:

$$\eta_{i,\text{final},k} = (1 - w_k) z_{c,k}(\eta_{i,\text{clinical},k}) + w_k z_{p,k}(\eta_{i,\text{PSG},k}) \,. \tag{8}$$

The weight w_k is selected from $\{0, 0.1, \ldots, 1\}$ by maximizing the validation Harrell C-index. The selected weight and both standardization transformations are fixed during held-out and external-cohort evaluation.

4 Experimental Setup

4.1 Cohorts, Incident Outcomes, and Data Splits

We used 26,782 participants from three Human Sleep Project (HSP) cohorts(Li et al., 2026), I0002, I0004, and I0006 for model development and held-out evaluation. An additional 1,000 participants from the Sleep Heart Health Study (SHHS) were reserved for external evaluation (Quan et al., 1997). Cohort characteristics and data partitions are summarized in Appendix Table S2. Endpoint-specific sample sizes and event counts vary according to baseline disease status and follow-up availability.

Available PSG channels were mapped to the six modalities defined in Section 3.1 and divided into 30-s windows. Recordings were truncated at 12 h, and channel- and window-level masks identified valid inputs. Missing modalities were not imputed. Sampling rates, channel mappings, and signal-quality criteria are provided in Appendix S1. Downstream features were standardized using statistics estimated from the corresponding training partition.

Participant-level partitions contained 19,073 training, 3,867 validation, and 3,842 test participants. For pooled evaluation, downstream models were fitted on the combined training partitions and selected on the combined validation partitions. For leave-one-cohort-out (LOCO) evaluation, the held-out cohort was excluded from downstream training, feature standardization, and model selection. External SHHS evaluation used models, normalization statistics, and fusion parameters derived from the development cohorts, without refitting on SHHS.

Incident outcomes are derived from ICD-9-CM(Centers for Medicare and Medicaid Services & National Center for Health Statistics, 2011) and ICD-10-CM(Centers for Medicare and Medicaid Services & National Center for Health Statistics, 2014)diagnoses in the linked Electronic Health Records (EHR). For each endpoint, we excluded participants diagnosed before or within 30 days after PSG and measured time from the day-30 landmark to the first qualifying diagnosis. Participants without post-landmark follow-up were also excluded. We evaluated ten outcomes: atrial fibrillation/flutter, heart failure, ischemic heart disease, acute kidney injury, acute stroke, cardiomyopathy, chronic kidney disease, COPD, dyslipidemia, and valvular heart disease. The same definitions were applied to 33 exploratory outcomes reported in Appendix S3

4.2 Comparisons and Evaluation

The Clinical baseline is a ridge-penalized Cox model using age and sex. To compare pretrained representations, we freeze our encoder and the public SleepFM encoder and apply the same Cox readout and model-selection procedure. SleepFM serves as the primary PSG representation baseline(Thapa et al., 2026).Ours Full PSG adds endpoint-specific whole-night aggregation and modality-specific risk heads to our frozen encoders. Ours PSG + Clinical combines its PSG score with the clinical score using the validation-selected fusion described in Section 3.4.

The Clinical baseline is a ridge-penalized Cox model using age and sex. For the matched frozen-probe comparison, our SSL encoders and the public SleepFM encoder are frozen and evaluated using the same Cox readout and model-selection procedure. This comparison measures how much prognostic information can be extracted from each pretrained representation under the same lightweight downstream model. SleepFM is the primary external representation baseline because it is a large-scale multimodal PSG model that has been evaluated for future disease prediction (Thapa et al., 2026).

Ours Full PSG uses the frozen modality encoders together with endpoint-specific whole-night aggregators and modality-specific risk heads. Available modality scores are combined at the risk level. Ours PSG + Clinical combines the resulting PSG score with the age-and-sex score using the validation-selected fusion method.

All comparisons use identical endpoint definitions, eligibility criteria, event times, censoring rules, and held-out participants. We report Harrell's C-index with 95% confidence intervals obtained from 1,000 participant-level bootstrap samples, stratified by cohort for pooled analyses. Differences between models are evaluated using paired bootstrap samples. Optimization settings and complete hyperparameter grids are provided in Appendix **??**.

Representation analyses. We measure similarity between frozen modality representations using linear CKA(Kornblith et al., 2019). We adjust for age, sex, and cohort and subtract a within-cohort permutation baseline, yielding permutation-adjusted CKA. Endpoint-specific modality information is assessed by comparing Cox models fitted with all modality blocks and with one block removed on the same participants. A dimension-matched PCA analysis is included as a sensitivity analysis. Full estimation details are provided in Appendix A.1.

Missing inputs and clinical complementarity. Missing-input robustness is evaluated by removing prespecified modalities at test time while keeping model parameters fixed. Each condition is compared with natural-input performance on the same participants. Clinical complementarity is assessed by comparing the age-and-sex model with the fused clinical–PSG model. For atrial fibrillation and heart failure, we additionally report Aalen–Johansen cumulative incidence curves using validation-fixed risk groups and treating recorded death as a competing event.

5 Results

5.1 Multi-Disease Risk Prediction

Table 1 natural-input performance across ten incident disease outcomes. Under the matched frozen Cox probe, SleepFM achieved a higher mean C-index than our frozen representation (0.692 vs 0.681). After endpoint-specific downstream modeling, Ours Full PSG reached 0.696 and produced higher point estimates than the SleepFM probe for eight of ten outcomes. Larger differences were observed for acute stroke (0.694 vs 0.672), chronic kidney disease (0.717 vs 0.692), and valvular disease (0.688 vs 0.673), whereas SleepFM remained higher for atrial fibrillation and ischemic heart disease.

Combining PSG with age and sex gave the highest point estimate for every outcome. The resulting mean C-index was 0.717, compared with 0.696 for Ours Full PSG, 0.692 for the SleepFM probe, and 0.672 for age and sex alone. Thus, the main gain arises from endpoint-specific use of the pretrained representations and their complementarity with demographic information, rather than uniformly stronger frozen features.

Six harmonized outcomes were also evaluated in the external cohort. PSG-only C-indices ranged from 0.51 to 0.69. The combined model performed best for atrial fibrillation (0.74; 95% CI, 0.70–0.79), followed by stroke (0.69; 0.64–0.74). Complete external results and the extended 33-endpoint analysis are reported in Appendix Table S3 and is not used to define the main comparison.

5.2 Shared Structure and Modality-Specific Predictive Information

Figure 2 compares cross-modal representation similarity with the predictive information added by each modality. Before disease-specific adaptation,cardiac and plethysmographic representations showed the strongest similarity, ($\Delta\mathrm{CKA} = 0.755$, 95% CI: 0.727–0.779), followed by neural and muscle representations, $\Delta\mathrm{CKA} = 0.218(0.172\text{–}0.272)$. Similarity nevertheless varied substantially across modality pairs.

After endpoint adaptation, the modalities contributed differently across outcomes. For atrial fibrillation, cardiac features produced the largest dimension-matched increment ($\Delta\mathrm{C} = 0.027$, 95% CI, 0.009–0.045). Using the native feature blocks, neural information produced the largest increment, $\Delta C = 0.026$ (0.009–0.043). Heart failure and ischemic heart disease showed different modality

Table 1: Natural-input prediction performance across primary incident outcomes.(Values are Harrell's C-indices with 95% confidence intervals.)

Outcome	N/ Events	Clinical	Ours frozen Cox	SleepFM frozen Cox	Ours PSG	Ours + Clinical
Atrial fibrillation/flutter	2,821/ 139	0.716 [0.668, 0.762]	0.677 [0.613, 0.734]	0.744 [0.687, 0.799]	0.713 [0.660, 0.765]	**0.749** [0.702, 0.791]
Heart failure	2,734/ 191	0.713 [0.676, 0.751]	0.710 [0.674, 0.749]	0.695 [0.652, 0.736]	0.704 [0.665, 0.745]	**0.727** [0.690, 0.767]
Ischemic heart disease	2,529/ 254	0.693 [0.662, 0.727]	0.670 [0.634, 0.705]	0.669 [0.635, 0.701]	0.656 [0.621, 0.693]	**0.699** [0.666, 0.734]
Acute kidney injury	2,927/ 206	0.684 [0.644, 0.720]	0.702 [0.668, 0.739]	0.733 [0.696, 0.771]	0.735 [0.701, 0.772]	**0.751** [0.718, 0.786]
Acute stroke	2,955/ 123	0.655 [0.604, 0.704]	0.673 [0.615, 0.725]	0.672 [0.622, 0.721]	0.694 [0.643, 0.741]	**0.711** [0.666, 0.753]
Cardiomyopathy	2,978/ 93	0.616 [0.562, 0.673]	0.692 [0.629, 0.756]	0.707 [0.650, 0.768]	0.709 [0.660, 0.765]	**0.713** [0.664, 0.769]
Chronic kidney disease	2,846/ 194	0.708 [0.668, 0.743]	0.694 [0.654, 0.732]	0.692 [0.656, 0.729]	0.717 [0.683, 0.755]	**0.747** [0.715, 0.781]
COPD	2,893/ 111	0.667 [0.616, 0.720]	0.672 [0.615, 0.725]	0.689 [0.637, 0.741]	0.697 [0.639, 0.751]	**0.721** [0.668, 0.771]
Dyslipidemia	1,691/ 358	0.630 [0.601, 0.662]	0.641 [0.612, 0.671]	0.644 [0.616, 0.673]	0.646 [0.617, 0.675]	**0.660** [0.633, 0.689]
Valvular heart disease	2,477/ 295	0.643 [0.610, 0.678]	0.675 [0.644, 0.706]	0.673 [0.644, 0.704]	0.688 [0.658, 0.719]	**0.694** [0.665, 0.725]
Ten-outcome mean	—	0.672 [0.654, 0.691]	0.681 [0.661, 0.699]	0.692 [0.674, 0.709]	0.696 [0.679, 0.713]	**0.717** [0.702, 0.733]

Table 2: Downstream fusion ablation under pooled and leave-one-cohort-out (LOCO) evaluation. (Values are C-indices with 95% confidence intervals.)

Downstream variant	Pooled	LOCO-I0002	LOCO-I0006	LOCO-I0004
Mean representation + shared Cox	0.690 [0.672, 0.706]	0.661 [0.650, 0.672]	0.621 [0.613, 0.629]	**0.665** [0.636, 0.692]
Concatenated representations + shared Cox	0.675 [0.657, 0.693]	0.653 [0.641, 0.663]	0.618 [0.610, 0.626]	0.657 [0.629, 0.684]
Modality-specific risk composition (MRC)	**0.696** [0.679, 0.713]	**0.670** [0.658, 0.681]	**0.624** [0.615, 0.631]	0.647 [0.622, 0.671]

rankings. Thus, cross-modal similarity does not imply equivalent predictive value or substitutability.

The controlled fusion results as shown in Table 2. Mean representation fusion and risk-level composition performed similarly, while concatenation generally yielded lower C-indices. Together, the representation and added-modality analyses suggest that related PSG modalities can provide complementary information for different outcomes, supporting the use of modality-specific prediction pathways.

5.3 Robustness to Missing PSG Inputs

Across the nine missing-input configurations in Figure 3, our model retained more of its natural-input discrimination than SleepFM in six. The largest difference occurred after jointly removing

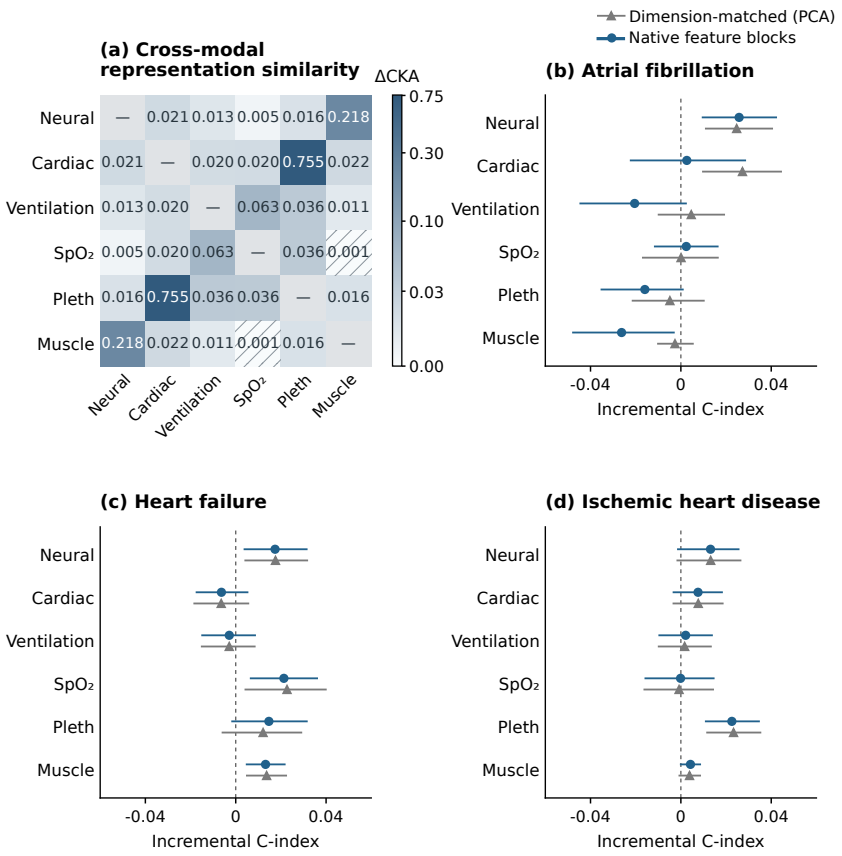

Figure 2: Shared and endpoint-specific information across PSG modalities. (a) Pairwise ΔCKA between frozen representations, adjusted for age, sex, and cohort and corrected by within-cohort permutation. Diagonal entries are undefined; hatching marks pairs with adjusted permutation-test $q \geq 0.05$. (b–d) Change in C-index from adding each modality for atrial fibrillation/flutter, heart failure, and ischemic heart disease. Circles use native features; triangles use PCA-matched 32-dimensional inputs. Bars show paired-bootstrap 95% confidence intervals.

neural, respiratory, and muscle signals: the ten-outcome mean C-index changed by (-0.021) for our model and (-0.070) for SleepFM, corresponding to a paired difference of $+0.049$ $(95\%\,\mathrm{CI}: 0.031\text{–}0.066)$. This setting leaves cardiac, SpO_2, and plethysmographic information available and shows that the proposed model can preserve much of its natural-input performance even after three physiological systems are removed.

Removing only respiratory or muscle signals had little effect on our model, with mean changes of (+0.002) and (-0.001), respectively. SleepFM showed smaller declines when removing SpO_2 alone (0.000 vs. -0.017) or jointly removing respiratory, SpO_2, and plethysmography (-0.015 versus -0.024). Thus, performance retention depends on which physiological systems are unavailable rather than only on the number of removed modalities.

5.4 Incremental Value Beyond Demographic Risk

Figure 4 evaluates whether PSG provides information beyond the age-and-sex baseline.

Adding PSG increased the point-estimate C-index for all eight outcomes included in this analysis, and the paired 95% confidence interval was above zero for six. The largest gains were observed for cardiomyopathy, $\Delta C = +0.113$ $(95\%\,\mathrm{CI}: 0.060\text{–}0.167)$, and type 2 diabetes, $\Delta C = +0.091$ $(0.044\text{–}0.142)$. Positive point estimates were also observed for AF, HF, IHD, acute stroke, CKD, and COPD.

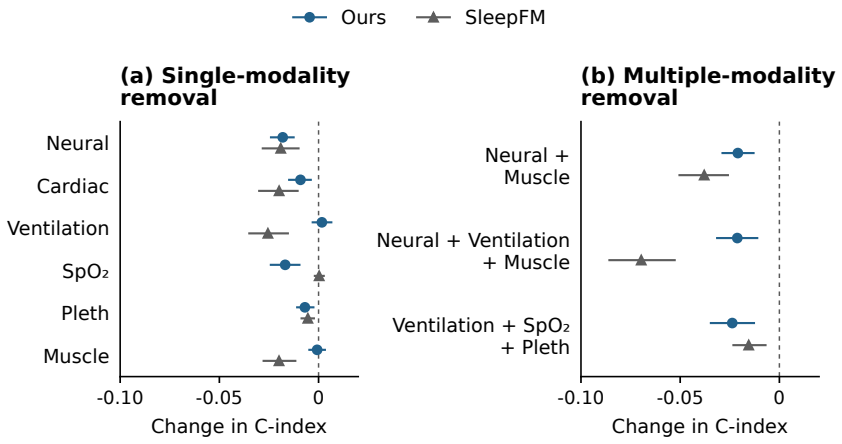

Figure 3: Sensitivity to missing PSG modalities across ten outcomes. Mean change in C-index from natural-input performance after (a) single- and (b) multiple-modality removal; higher values indicate better retention. Ours uses 30% modality dropout, whereas SleepFM uses frozen features with ridge Cox regression. Bars show 95% CIs from 1,000 cohort-stratified participant bootstraps.

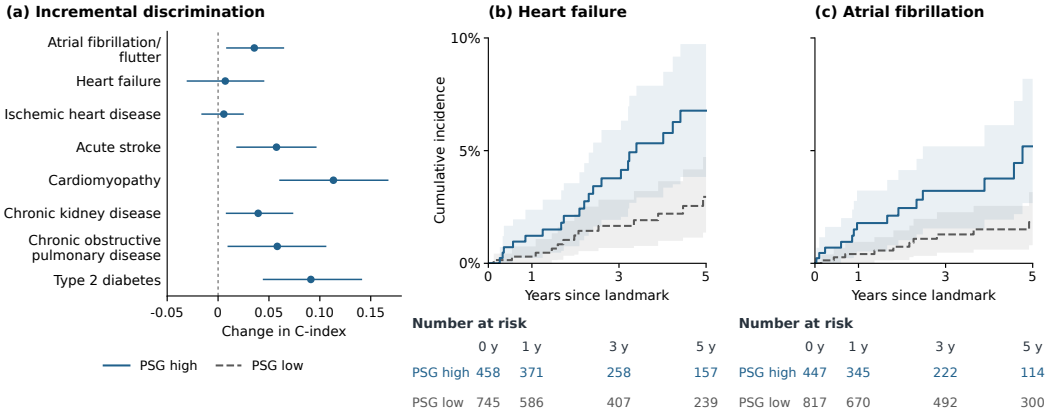

Figure 4: Complementarity of PSG and demographic risk. (a) Change in C-index after adding PSG to age and sex. (b,c) Cumulative incidence of heart failure and atrial fibrillation/flutter among participants below the validation-defined demographic-risk threshold, stratified by PSG risk. Shading and bars show 95% CIs; numbers at risk are shown below.

The cumulative-incidence analysis provides a more direct view of the same effect. Among participants below the validation-defined age-and-sex risk threshold, five-year HF incidence was 6.8% in the PSG-high group and 2.9% in the PSG-low group. For AF/flutter, the corresponding values were 5.2% and 1.8%. Adjusted hazard-ratio analyses using additional recorded clinical covariates are reported in Appendix Figure 5. They are treated as sensitivity analyses rather than as separate disease-specific clinical models.

Thus, PSG further separates future risk among individuals placed in the same lower demographic-risk group. This result complements the C-index analysis: the PSG score adds information not captured by age and sex alone, and this additional information is visible in subsequent disease incidence.

6 DISCUSSION

Our results show that PSG modalities share participant-level structure but are not interchangeable for disease-risk prediction. Although cardiac and plethysmographic representations showed strong similarity, the information added by each modality varied across endpoints. This finding supports using cross-modal supervision during pretraining while retaining modality-specific downstream pathways.

The matched-probe comparison further separates representation quality from downstream modeling. SleepFM performed better under the common frozen Cox probe, whereas endpoint-specific aggregation and risk estimation improved the use of our representations and produced higher point estimates for most outcomes. Separate modality pathways also permit prediction from incomplete inputs, although performance depends on which physiological systems are unavailable. Combining PSG with age and sex further improved discrimination, indicating that overnight physiology provides information complementary to this demographic baseline.

Several boundaries remain. LOCO experiments measure downstream transfer because held-out cohorts may have contributed unlabeled data during pretraining. EHR-derived outcomes are subject to variation in coding and follow-up, and age and sex form a deliberately limited clinical comparator. Future work should evaluate cohorts excluded from both pretraining and downstream development, adjudicated outcomes, stronger clinical baselines, and calibrated absolute risk. Overall, our findings support separating cross-modal representation learning from downstream risk integration when modeling heterogeneous physiological signals.

AI USE STATEMENT

In this work, we used generative AI tools such as Codex to assist in grammar correction, language polishing, structure organization, formula formatting, and some data analysis tasks. The latter included debugging analysis code, discussing result analysis and its interpretation, etc.

The author reviewed the text generated by AI assistance and checked the analysis code, verified the statistical methods, and compared the reported values with the experimental results for validation.

All methodological decisions and final interpretations were the responsibility of the author. The author is fully responsible for the entire content of the paper, including all the text generated by AI assistance, the analysis results, and related achievements.

ETHICS STATEMENT

This study analyzes de-identified PSG and linked clinical data under the applicable data-use agreements. The model is intended for research and has not been validated for clinical decision-making. Dataset biases and variation across populations may limit generalizability.

REPRODUCIBILITY STATEMENT

The Methods and Experimental Setup describe the model, outcome definitions, data splits, and evaluation procedures. Additional preprocessing details, training settings, and hyperparameter grids are provided in the Appendix.

AUTHOR CONTRIBUTIONS

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

A. Azarbarzin, S. A. Sands, K. L. Stone, L. Taranto-Montemurro, L. Messineo, P. I. Terrill, S. Ancoli-Israel, K. Ensrud, S. Purcell, D. P. White, S. Redline, and A. Wellman. The hypoxic burden of sleep apnoea predicts cardiovascular disease-related mortality: The Osteoporotic Fractures

in Men Study and the Sleep Heart Health Study. *European Heart Journal*, 40(14):1149–1157, 2019. doi: 10.1093/eurheartj/ehy624. URL https://doi.org/10.1093/eurheartj/ehy624.

A. Azarbarzin, S. A. Sands, L. Taranto-Montemurro, D. Vena, T. Sofer, S.-W. Kim, K. L. Stone, D. P. White, A. Wellman, and S. Redline. The sleep apnea-specific hypoxic burden predicts incident heart failure. *Chest*, 158(2):739–750, 2020. doi: 10.1016/j.chest.2020.03.053. URL https://doi.org/10.1016/j.chest.2020.03.053.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1298–1312. PMLR, 2022. URL https://proceedings.mlr.press/v162/baevski22a.html.

Centers for Medicare and Medicaid Services and National Center for Health Statistics. ICD-9-CM official guidelines for coding and reporting. Technical report, U.S. Department of Health and Human Services, 2011. URL https://archive.cdc.gov/#/details?q=ICD%20&start=0&rows=10&url=https://www.cdc.gov/nchs/data/icd/icd9cm_guidelines_2011.pdf.

Centers for Medicare and Medicaid Services and National Center for Health Statistics. ICD-10-CM official guidelines for coding and reporting. Technical report, U.S. Department of Health and Human Services, 2014. URL https://archive.cdc.gov/#/details?q=ICD%20&start=0&rows=10&url=https://www.cdc.gov/nchs/data/icd/icd10cm_guidelines_2014.pdf.

Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=B1xwcyHFDr.

V. K. Kapur, D. H. Auckley, S. Chowdhuri, D. C. Kuhlmann, R. Mehra, K. Ramar, and C. G. Harrod. Clinical practice guideline for diagnostic testing for adult obstructive sleep apnea: An American Academy of Sleep Medicine clinical practice guideline. *Journal of Clinical Sleep Medicine*, 13(3):479–504, 2017. doi: 10.5664/jcsm.6506. URL https://doi.org/10.5664/jcsm.6506.

Dani Kiyasseh, Tingting Zhu, and David A. Clifton. CLOCS: Contrastive learning of cardiac signals across space, time, and patients. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5606–5615. PMLR, 2021. URL https://proceedings.mlr.press/v139/kiyasseh21a.html.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Qichen Li, Shenghan Wen, Haoqi Sun, Wolfgang Ganglberger, Ayush Tripathi, Niels Turley, Samuel Waters, Arnav Gupta, Aditya Gupta, Manohar Ghanta, Bruce Nearing, Han Wu, Katie L. Stone, Chad Robichaux, Zhiyong Zhang, Qiao Li, Gauri Ganjoo, Christine Tsien Silvers, Bharath Gunapati, Kiran Maski, Samaneh Nasiri, Dennis Hwang, Lynn Marie Trotti, Umakanth Katwa, Gari D. Clifford, Emmanuel Mignot, Robert J. Thomas, and M. Brandon Westover. The Human Sleep Project. Brain Data Science Platform, 2026. URL https://doi.org/10.60508/m3sw-rz13. Version 3.0.

L. Paixao, P. Sikka, H. Sun, A. Jain, J. Hogan, R. Thomas, and M. B. Westover. Excess brain age in the sleep electroencephalogram predicts reduced life expectancy. *Neurobiology of Aging*, 88:150–155, 2020. doi: 10.1016/j.neurobiolaging.2019.12.015. URL https://doi.org/10.1016/j.neurobiolaging.2019.12.015.

S. F. Quan, B. V. Howard, C. Iber, J. P. Kiley, F. J. Nieto, G. T. O'Connor, D. M. Rapoport, S. Redline, J. Robbins, J. M. Samet, and P. W. Wahl. The Sleep Heart Health Study: Design, rationale, and methods. *Sleep*, 20(12):1077–1085, 1997. doi: 10.1093/sleep/20.12.1077. URL https://doi.org/10.1093/sleep/20.12.1077.

Rahul Thapa, Bryan He, Magnus Ruud Kjaer, Hyatt Moore, IV, Gauri Ganjoo, Emmanuel Mignot, and James Zou. SleepFM: Multi-modal representation learning for sleep across brain activity, ECG and respiratory signals. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48019–48037. PMLR, 2024. URL https://proceedings.mlr.press/v235/thapa24a.html.

Rahul Thapa, Magnus Ruud Kjaer, Bryan He, Ian Covert, Hyatt Moore, IV, Umaer Hanif, Gauri Ganjoo, M. Brandon Westover, Poul Jennum, Andreas Brink-Kjaer, Emmanuel Mignot, and James Zou. A multimodal sleep foundation model for disease prediction. *Nature Medicine*, 32(2):752–762, 2026. doi: 10.1038/s41591-025-04133-4. URL https://doi.org/10.1038/s41591-025-04133-4.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision – ECCV 2020*, pp. 776–794. Springer, 2020a. doi: 10.1007/978-3-030-58621-8_45. URL https://doi.org/10.1007/978-3-030-58621-8_45.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *Advances in Neural Information Processing Systems*, volume 33, pp. 6827–6839. Curran Associates, Inc., 2020b. URL https://proceedings.neurips.cc/paper/2020/hash/4c2e5eaae9152079b9e95845750bb9ab-Abstract.html.

Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/1102a326d5f7c9e04fc3c89d0ede88c9-Abstract.html.

L. Zhang, H. Wu, X. Zhang, X. Wei, F. Hou, and Y. Ma. Sleep heart rate variability assists the automatic prediction of long-term cardiovascular outcomes. *Sleep Medicine*, 67:217–224, 2020. doi: 10.1016/j.sleep.2019.11.1259. URL https://doi.org/10.1016/j.sleep.2019.11.1259.

A Appendix

This appendix provides additional details on signal processing, training configurations, and representation analyses.

A.1 Representation Analysis Details

CKA analysis. For each modality, we constructed a nightly representation using the masked mean and standard deviation of its valid frozen-encoder tokens. We adjusted these representations for age, sex, and cohort by fitting a separate linear nuisance model for each modality on training and validation participants with that modality observed, then applying the fitted model to the held-out test set. Debiased linear CKA was computed on participants with both modalities observed. The primary effect, $\Delta\mathrm{CKA}$, was defined as the adjusted paired CKA minus the mean null CKA obtained from 1,000 permutations of participant correspondence within cohort. Permutation p-values were adjusted for false discovery rate (FDR) across the 15 modality pairs. Raw paired CKA without covariate adjustment and covariate-adjusted paired CKA are reported separately in the supplementary results.

PCA sensitivity analysis. PCA was not used in the CKA analysis. In the separate added-modality experiment, the all-modality and other-modality feature sets were each reduced to 32 total dimensions, rather than 32 dimensions per modality. Standardization and PCA were fitted separately for each feature set using the corresponding training subset and applied unchanged to validation and test data. A separate ridge-penalized Cox probe was fitted for each input, with regularization selected on validation data.

Table S1: Signal types, canonical channel names, and target sampling rates. The input specification covers six physiological modalities and up to 18 channels. Sampling rates are target rates rather than necessarily the original acquisition rates.

Physiological modality	Signal type	Canonical channel name(s)	Maximum channels	Target rate (Hz)
Neural	Electroencephalography (EEG)	F3, F4, C3, C4, O1, O2	6	128
Neural	Electrooculography (EOG)	E1, E2	2	128
Cardiac	Electrocardiography (ECG)	ECG	1	128
Ventilation	Airflow	airflow	1	16
Ventilation	Thoracic respiratory effort	thorax	1	16
Ventilation	Abdominal respiratory effort	abdomen	1	16
Ventilation	Snoring-related signal	snore	1	16
Oxygenation	Peripheral oxygen saturation	SpO2	1	4
Plethysmography	Pulse plethysmographic waveform	pleth	1	64
Muscle	Chin electromyography (EMG)	chin	1	128
Muscle	Leg electromyography (EMG)	left_leg, right_leg	2	128
Total	Six physiological modalities	—	18	Modality-specific

Table S2: Cohort characteristics and participant-level data splits. Age is reported in years as mean ± standard deviation. Demographic statistics describe the full analysis sample within each cohort.

(a) Participant counts and splits

Cohort	Total	Training	Validation	Internal test	External test
I0002	11,763	8,236	1,764	1,763	—
I0006	11,109	7,762	1,704	1,643	—
I0004	3,910	3,075	399	436	—
Development total	26,782	19,073	3,867	3,842	—
SHHS1	1,000	—	—	—	1,000
All cohorts	27,782	19,073	3,867	3,842	1,000

(b) Demographic characteristics

Cohort	Age, mean ± SD	Male, n (%)
I0002	56.2 ± 16.9	6,435 (54.7%)
I0006	54.6 ± 15.3	5,373 (48.4%)
I0004	51.4 ± 18.9	2,307 (59.0%)
Development total	54.8 ± 16.6	14,115 (52.7%)
SHHS1	67.3 ± 13.4	447 (44.7%)
All cohorts	55.3 ± 16.7	14,562 (52.4%)

Table S3: Prediction performance across 33 incident disease outcomes. Values are Harrell's C-indices with 95% confidence intervals. Clinical denotes the age-and-sex risk score. N and Events indicate the endpoint-specific test sample size and event count. An asterisk marks outcomes with fewer than 50 observed events. COPD denotes chronic obstructive pulmonary disease.

Outcome	N/Events	Age + sex	Ours PSG	Ours PSG + Clinical
Atrial fibrillation/flutter	2,821/139	0.716 [0.668, 0.762]	0.713 [0.660, 0.765]	0.749 [0.702, 0.791]
Heart failure	2,734/191	0.713 [0.676, 0.751]	0.704 [0.665, 0.745]	0.727 [0.690, 0.767]
Ischemic heart disease	2,529/254	0.693 [0.662, 0.727]	0.656 [0.621, 0.693]	0.699 [0.666, 0.734]
Acute kidney injury	2,927/206	0.684 [0.644, 0.720]	0.735 [0.701, 0.772]	0.751 [0.718, 0.786]
Acute stroke	2,955/123	0.655 [0.604, 0.704]	0.694 [0.643, 0.741]	0.711 [0.666, 0.753]
Cardiomyopathy	2,978/93	0.616 [0.562, 0.673]	0.709 [0.660, 0.765]	0.713 [0.664, 0.769]
Chronic kidney disease	2,846/194	0.708 [0.668, 0.743]	0.717 [0.683, 0.755]	0.747 [0.715, 0.781]
COPD	2,893/111	0.667 [0.616, 0.720]	0.697 [0.639, 0.751]	0.721 [0.668, 0.771]
Dyslipidemia	1,691/358	0.630 [0.601, 0.662]	0.646 [0.617, 0.675]	0.660 [0.633, 0.689]
Valvular heart disease	2,477/295	0.643 [0.610, 0.678]	0.688 [0.658, 0.719]	0.694 [0.665, 0.725]
Anxiety disorder	2,453/359	0.610 [0.580, 0.640]	0.577 [0.547, 0.609]	0.620 [0.590, 0.651]
Asthma	2,615/145	0.584 [0.535, 0.629]	0.592 [0.541, 0.641]	0.611 [0.560, 0.656]
Bipolar disorder*	3,079/32	0.580 [0.483, 0.677]	0.604 [0.497, 0.705]	0.610 [0.502, 0.709]

Continued on next page

Table S4: Model configuration and training hyperparameters.

Component	Setting
Input and encoders	30-s windows; 10 windows per 5-min downstream token; up to 144 tokens/night. Six independently parameterized encoders with representation width 128.
SSL objectives	Mask ratio 0.50; Smooth L1 latent-prediction loss; cross-modal loss weight 0.20; context-modality dropout 0.35.
EMA teacher	Decay scheduled from 0.996 to 0.9999; no teacher gradients.
SSL optimization	AdamW; learning rate 5×10^{-5}; weight decay 0.05; 3 epochs, followed by 1 continuation epoch on repaired data at 3×10^{-5}.
Downstream training	Frozen encoders; AdamW; initial learning rate 0.002; batch size 48; up to 45 epochs; early-stopping patience 8; random seed 42.
I0004 training expansion	Warm-start downstream heads; learning rate 5×10^{-4}; up to 20 additional epochs; early-stopping patience 5.
Risk objective	Endpoint-specific Cox loss; unimodal auxiliary loss weight 0.20; no synthetic downstream modality dropout in this configuration.
Clinical fusion	Age and sex; PSG fusion weight selected from $\{0, 0.1, \ldots, 1.0\}$ by maximizing source-validation Harrell's C-index.

Table S3 continued

Outcome	N/Events	Age + sex	Ours PSG	Ours PSG + Clinical
Cerebrovascular disease (broad)	2,719/203	0.700 [0.663, 0.734]	0.669 [0.629, 0.704]	0.703 [0.668, 0.734]
Dementia/Alzheimer disease	3,127/84	0.801 [0.746, 0.851]	0.701 [0.636, 0.765]	0.791 [0.740, 0.839]
Depression	2,328/271	0.576 [0.538, 0.610]	0.567 [0.526, 0.603]	0.606 [0.571, 0.638]
Epilepsy	3,007/53	0.509 [0.422, 0.594]	0.567 [0.471, 0.658]	0.567 [0.471, 0.658]
Fatty liver disease	2,960/183	0.518 [0.477, 0.564]	0.565 [0.519, 0.610]	0.565 [0.519, 0.610]
Hemorrhagic stroke*	3,131/37	0.599 [0.491, 0.704]	0.641 [0.551, 0.728]	0.631 [0.528, 0.730]
Hypertension	1,409/248	0.645 [0.609, 0.681]	0.647 [0.613, 0.685]	0.671 [0.637, 0.704]
Hyperthyroidism*	3,109/21	0.657 [0.530, 0.764]	0.674 [0.550, 0.784]	0.657 [0.530, 0.764]
Hypothyroidism	2,719/125	0.654 [0.600, 0.704]	0.590 [0.540, 0.649]	0.653 [0.602, 0.707]
Ischemic stroke (broad)	2,898/139	0.682 [0.639, 0.722]	0.674 [0.624, 0.720]	0.694 [0.647, 0.736]
Metabolic syndrome*	3,118/38	0.534 [0.441, 0.623]	0.549 [0.449, 0.631]	0.549 [0.449, 0.631]
Myocardial infarction	3,070/69	0.667 [0.602, 0.731]	0.658 [0.583, 0.724]	0.658 [0.583, 0.724]
Obesity	1,893/329	0.533 [0.501, 0.565]	0.649 [0.619, 0.679]	0.651 [0.620, 0.681]
Parkinson disease*	3,141/31	0.831 [0.756, 0.892]	0.779 [0.708, 0.850]	0.846 [0.777, 0.905]
Peripheral artery disease	2,801/213	0.679 [0.644, 0.714]	0.636 [0.595, 0.675]	0.686 [0.650, 0.721]
Posttraumatic stress disorder*	3,121/44	0.640 [0.551, 0.726]	0.649 [0.566, 0.729]	0.640 [0.551, 0.726]
Substance use disorder	2,770/120	0.523 [0.471, 0.574]	0.579 [0.524, 0.635]	0.574 [0.514, 0.634]
Type 1 diabetes*	3,052/32	0.575 [0.477, 0.680]	0.769 [0.681, 0.843]	0.769 [0.681, 0.843]
Type 2 diabetes	2,441/246	0.567 [0.529, 0.604]	0.656 [0.619, 0.694]	0.664 [0.627, 0.700]
Venous thromboembolism	2,977/135	0.663 [0.615, 0.710]	0.675 [0.624, 0.730]	0.697 [0.644, 0.748]

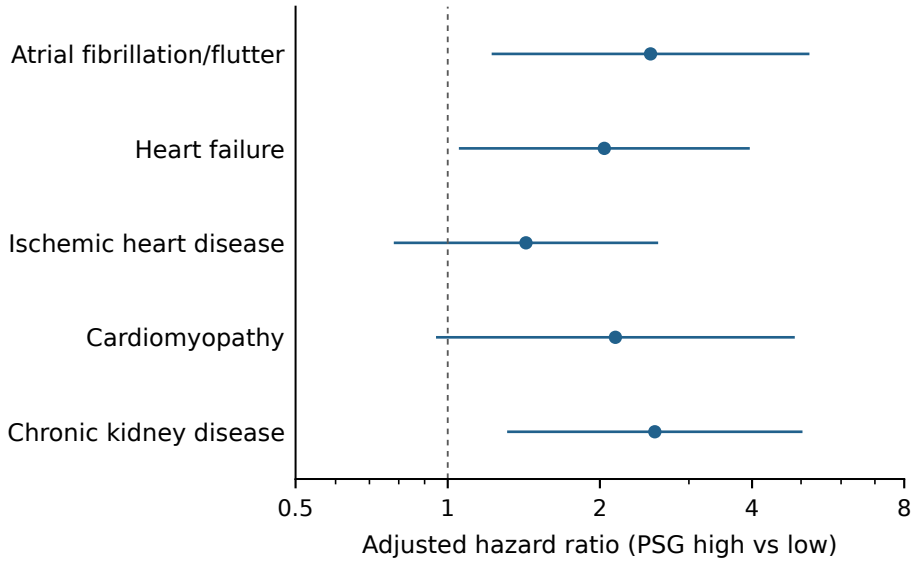

Figure 5: Adjusted cause-specific hazard ratios for PSG high versus low among participants whose age-and-sex risk score was below the endpoint-specific source-validation median. PSG groups were split at the corresponding validation-median PSG score, with no middle group excluded; follow-up began 30 days after PSG. Cohort-stratified Cox models adjusted for age, sex, recorded pre-PSG smoking, and diabetes. Points and bars show hazard ratios and Wald 95% confidence intervals for atrial fibrillation/flutter, heart failure, ischemic heart disease, cardiomyopathy, and chronic kidney disease (32, 40, 50, 25, and 40 events, respectively).