BCG-FM: A FOUNDATION MODEL FOR AMBIENT CARDIAC HEALTH SENSING

Anonymous authors
Paper under double-blind review

ABSTRACT

Foundation models for wearable biosignals have matched or exceeded supervised specialists across a range of clinical tasks, yet all rely on modalities that require deliberate user action–wearing a device or visiting a sleep lab. We introduce BCG-FM, the first foundation model for ambient mechanical biosignals. A piezoelectric sensor embedded in the bed surface records ballistocardiography (BCG) each night without user effort; from a corpus of 2.75 million hours of nightly recordings from 145,857 individuals, we pretrain with participant-level contrastive learning. Frozen BCG-FM embeddings achieve 3.26-year MAE on biological-age estimation (the lowest reported for any ambient, contactless modality) and yield clinically relevant discrimination across 15 self-reported health conditions and two independent external clinical cohorts. Pretrained representations from only 500 labeled participants outperform a fully supervised baseline trained on 3,372, and representation quality scales log-linearly with contrastive batch size. A second pretraining stage, in which each minute predicts how the embeddings of neighbouring minutes differ from its own, further enriches the embeddings with sleep-stage, apnea, and heart-rate information while retaining age information. These results establish ambient, longitudinal mechanical biosignals as a viable modality for health foundation models.

1 INTRODUCTION

Foundation models pretrained on photoplethysmography (PPG) (Abbaspourazad et al., 2024; Pillai et al., 2025), electrocardiography (ECG) (Li et al., 2025; Coppola et al., 2024), and polysomnography (PSG) (Thapa et al., 2026) now rival or exceed supervised specialists on tasks from arrhythmia detection to biological age estimation. Yet these advances share a common constraint: they require deliberate user action: wearing a watch, attaching electrodes, or sleeping in a clinical lab.

Ballistocardiography (BCG) offers a fundamentally different paradigm. BCG is the whole-body mechanical recoil from cardiac ejection, respiration and movement, a bed-referenced force (unlike limb-worn accelerometry) with mechanical rather than electrical or optical transduction (unlike ECG/PPG). A piezoelectric sensor band embedded in the bed surface records this recoil with every heartbeat and breath, generating a continuous physiological record each night without user effort. Over 100,000 individuals already generate BCG data nightly through commercial sleep products, yet BCG remains absent from the foundation-model literature, limited to small-cohort supervised studies (Inan et al., 2015).

We present BCG-FM, the first foundation model for ambient mechanical biosignals. BCG-FM is trained and evaluated with participant-level contrastive learning on 145,857 users and 2.75 million hours of nightly BCG, the largest BCG corpus to date and orders of magnitude beyond prior BCG studies. The ambient, continuous nature of bed-based recording yields complete overnight physiological records without battery constraints or adherence requirements, enabling session-level modeling strategies that are infeasible for wearable modalities.

Frozen BCG-FM representations achieve **3.26**-year mean absolute error (MAE) on biological age estimation, the lowest reported for any ambient, contactless modality, and predict health conditions from self-reported labels, with clinically relevant performance on diabetes (AUROC 0.853), sleep apnea (0.792), and heart conditions (0.734). We validate generalization on two independent external clinical datasets, including heart failure detection (AUROC 0.841) on a held-out hospital cohort.

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.

These results demonstrate that the mechanical cardiac signal captured by an ambient bed sensor encodes rich, clinically actionable health information. Our contributions include:

(1) First ambient biosignal foundation model at scale. We pretrain on 136,575 participants (2.04M hours) from a commercially deployed bed sensor requiring no user action.
(2) Accurate estimation of age and disease. Frozen embeddings estimate biological age with an MAE of 3.26 years and predict diabetes (AUROC, 0.852), heart failure (0.822) and atrial fibrillation (0.954) across internal and external cohorts.
(3) Label-free enrichment of a pretrained foundation model. After optimizing contrastive pretraining, a second stage adds sleep-stage, apnea and heart-rate information (-14% heart-rate error) while retaining age information (§4.2, §5.6).
(4) Release of the Twilight dataset. We will share Twilight, 195 nights (1,944 hours) of bed BCG with simultaneous PSG and 1-minute sleep stages, with fixed splits and baselines (§3.3, Appendix D).

2 Related Work

2.1 Foundation models for wearable biosignals

The PPG foundation model of Abbaspourazad et al. (2024) pretrained a 1D EfficientNet encoder on $\sim$141K Apple Watch users via contrastive learning with participant-level pairing. Miller et al. (2025) extended this framework to 213K participants, defining the PpgAge biomarker on both a self-reported healthy cohort and a general population.[1] PaPaGei (Pillai et al., 2025) combines contrastive and generative objectives on 57K hours of PPG across 20 tasks; AnyPPG (Nie et al., 2025b) uses cross-modal PPG–ECG alignment on 100K hours from 58K participants for phenome-wide health profiling. Nie et al. (2025a) train a supervised CNN on 212K UK Biobank participants, linking predicted age acceleration to cardiovascular events. BCG-FM adopts the contrastive framework and encoder architecture of Abbaspourazad et al. (2024); the key differences are input modality (optical pulse vs. mechanical vibration), acquisition setting (wrist-worn vs. ambient bed), and our finding that PPG augmentation strategies do not transfer to BCG (Appendix A).

Beyond PPG, LSM (Narayanswamy et al., 2025) pretrains on up to 40 million hours of multimodal wearable data from 165K participants; SensorLM (Zhang et al., 2025) extends this with sensor–language alignment on 60 million hours. The Accelerometer Foundation Model (Abbaspourazad et al., 2025) uses knowledge distillation for wrist accelerometry. General-purpose time-series foundation models (e.g., Chronos, TimesFM) target forecasting rather than health representation learning and are not directly comparable. All require the user to wear a device.

2.2 Foundation models for clinical biosignals

ECGFounder (Li et al., 2025) pretrained on 10 million 12-lead ECGs from 1.8 million subjects with supervised multi-task learning across 150 diagnostic categories. HuBERT-ECG (Coppola et al., 2024) adapted masked prediction to 9.1 million ECGs across 164 conditions, and CSFM (Gu et al., 2026) pretrains a multimodal cardiac model on data from 1.7 million individuals. Both exploit rich multi-lead morphology but require skin-contact clinical electrodes.

These ECG models represent an upper bound given high-fidelity multi-lead input. BCG inevitably loses morphological detail but compensates with longitudinal scale (hundreds of nights per user vs. typically one clinical ECG) and ambient acquisition, enabling population-level pretraining infeasible in clinical settings.

2.3 Sleep and multi-modal foundation models

SleepFM (Thapa et al., 2026), the most directly comparable prior work, pretrains on PSG from $\sim$65K participants across multiple clinical cohorts, combining EEG, EOG, EMG, ECG, and respiratory channels with a leave-one-out contrastive objective resilient to missing channels. It evaluates on 130 conditions and biological age estimation.

[1] PpgAge's 2.43-yr MAE uses a self-reported healthy cohort (n=6,728); the general-population MAE of 3.2 yr is the more comparable benchmark given BCG-FM's HSS-based cohort.

Table 1: Dataset summary. **Pretrain**: corpus for self-supervised learning. **Healthy**: subset selected via Health Status Score (Section 3.4) for age estimation. **Disease**: pretrain subset with self-reported condition labels. **External 1/2**: independent clinical datasets not used for pretraining. F% = Female%

Split	Participants	Sessions	Segments	Hours	Sess/user	Age	BMI	F %
Pretrain	136 575	498 409	122M†	2 041 288	3.65	41.8	25.3	37.4
Disease (Train)	50 128	270 538	168M	2 793 434	5.40	41.8	25.3	39.8
Disease (Test)	12 532	67 817	42M	697 698	5.41	41.7	25.3	40.0
Healthy (Train)	3 372	20 172	15.7M	261 241	5.98	42.2	24.7	26.5
Healthy (Val)	1 071	6 636	4.7M	77 503	6.20	41.3	24.8	31.3
Healthy (Test)	4 708	28 247	22.2M	369 595	6.00	40.9	24.7	26.4
External 1 (LOO)	85	153	153	2.5	1.80	46.5	23.8	23.5
External 2 (LOO)	46	46	28 917	482	1.00	74.8	20.7	43.5
Total	**145 857**	**553 663**	—	**2 750 111**	3.80	41.8	25.3	36.8

†122M of 273M total segments used for pretraining after signal-quality filtering.

2.4 Biological age estimation from biosignals

Biological age estimation has emerged as a key benchmark for biosignal foundation models, spanning wrist-worn PPG (Miller et al., 2025), clinical ECG (Li et al., 2025; Lima et al., 2021), and multi-channel PSG (Brink-Kjær et al., 2022; Thapa et al., 2026). Performance varies widely with modality fidelity and cohort composition. We present a cross-modality comparison in Table 3 (§5.1); direct comparison requires caution given differences in test-set age distributions and health definitions.

2.5 Ballistocardiography

Ballistocardiography, the measurement of body recoil forces from cardiac ejection, was first described by Gordon (1877) and has seen renewed interest with unobtrusive sensors in beds, chairs, and scales (Inan et al., 2015). Prior ML work has focused on narrow tasks with limited cohorts: heart-rate estimation (Brüser et al., 2013), sleep staging (Yi et al., 2019), and heart- and respiration-rate detection (Albukhari et al., 2019), typically involving tens to hundreds of participants. BCG-FM is, to our knowledge, the first self-supervised model at foundation-model scale (>100K participants) for BCG, and the first to show that BCG supports general-purpose health prediction.

3 Data

3.1 Bed-based BCG recordings

BCG-FM is trained on nightly recordings from a commercially deployed bed-based sensor (Figure 1a). The sensor uses a piezoelectric strip embedded beneath the bed surface to capture the mechanical recoil of the body produced by cardiac ejection, respiration, and gross body movement. The raw signal is recorded at 500 Hz and downsampled to $f_s = 64$ Hz for model input, then transmitted to cloud storage, where it is de-identified in accordance with the data privacy policy and informed user consent. Recordings span the full sleep period (typically 5–10 hours) for each night the user sleep on their bed, yielding a dense physiological time series that requires no user action after initial set-up.

For pretraining, each night is segmented into non-overlapping 60-second windows. Each window is decomposed into two physiologically motivated channels: a *breathing* channel (0.5 s moving-average of the raw signal) and a *heart* channel (the residual after subtraction), giving the encoder explicit access to respiratory and cardiac timescales. Both channels are z-scored per window. Windows with excessive motion artifact or signal dropout are excluded via a signal-quality filter.

3.2 External clinical datasets

We validate BCG-FM on two independent clinical datasets not seen during pretraining or model selection. **External 1** (N=85, 153 nights): expert-annotated cardiac rhythm diagnoses (atrial

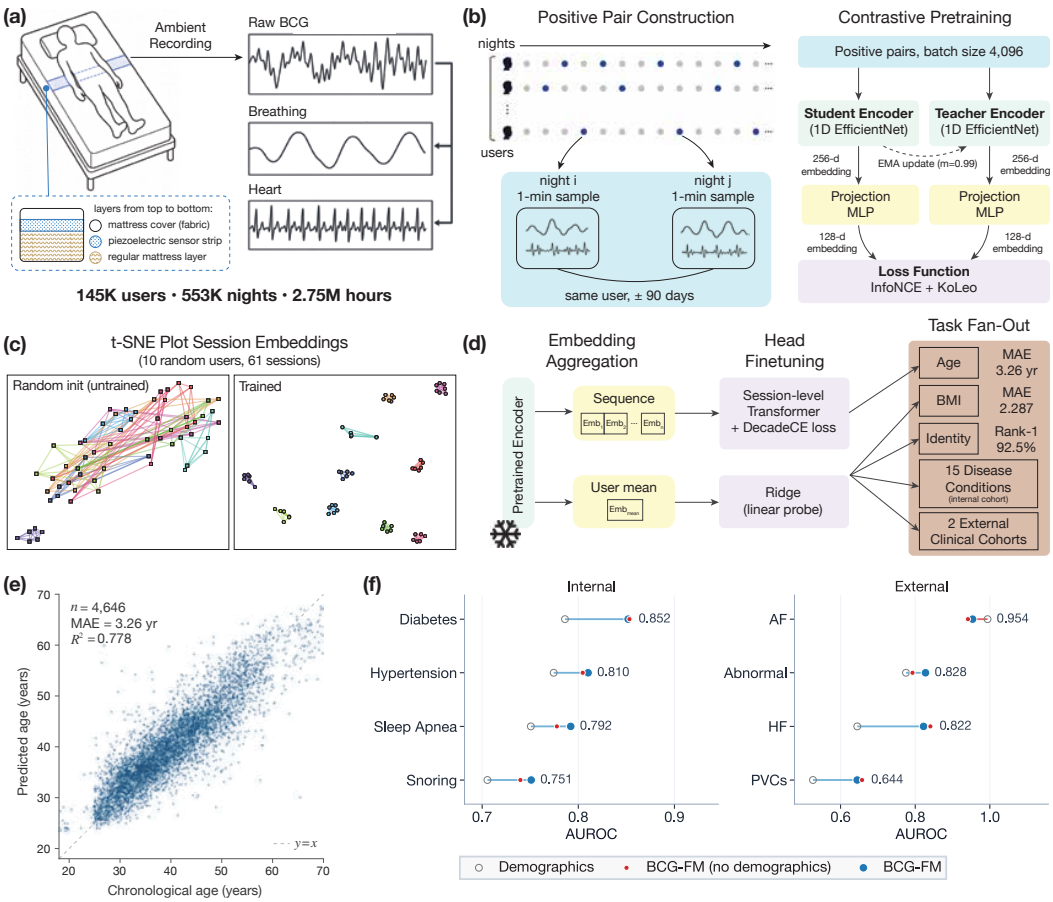

Figure 1: Overview of BCG-FM. (a) A piezoelectric bed sensor captures nightly BCG recordings, decomposed into breathing and heart channels. (b) Positive pairs from different nights of the same user (±90 days) train a dual-encoder via InfoNCE + KoLeo contrastive loss without augmentations; blue points mark the matched positive pairs in a batch, grey points the unmatched (negative) samples. (c) t-SNE of session embeddings shows clear per-user clustering after pretraining. (d) Frozen backbone embeddings are evaluated through a session-level Transformer for age and a Ridge linear probe for remaining tasks. (e) Biological age prediction (18–70 years, MAE 3.26 yr, coefficient of determination R^2=0.778). (f) Disease prediction AUROC on internal conditions and two independent external clinical cohorts.

fibrillation, AF; heart failure, HF; premature atrial/ventricular contractions, PACs/PVCs; sinus-normal); mean age 46.5 (Zhan et al., 2025). **External 2** (N=46, 28,917 windows): elderly cohort (mean age 74.8) with and without AF (Qiu et al., 2025). Both use subject-level leave-one-out cross validation.

3.3 Internal BCG–PSG Dataset

Twilight is a distinct internal dataset of 195 in-lab nights (1,944 hours) of bed BCG with simultaneous polysomnography, giving 1-minute sleep-stage, respiratory-event, and ECG heart-rate references on 10 lab beds not used in pretraining (Appendix D).

3.4 Labels and Downstream Targets

BCG-FM is evaluated on five task families: (1) **Demographics** (age, BMI); (2) **Identity retrieval** (Rank-1, MRR via cosine similarity); (3) **Disease prediction (internal)**: 15 self-reported condi-

tions; (4) **Disease prediction (external)**: expert-adjudicated cardiac diagnoses; (5) **Healthy cohort definition** for age estimation via Health Status Score (HSS).

Health Status Score (HSS). We define a Health Status Score integrating self-reported survey data (conditions, sleep quality, lifestyle, BMI; 0–100 pts), device-measured sleep architecture, and linked exercise metrics. Users are assigned to tiers via a grade $\times$ data-confidence matrix; Tier 1 ("healthy") yields 9,151 users for age estimation, less restrictive than PpgAge's self-reported healthy cohort but substantially more selective than a general-population approach. Full scoring details and tier distributions are in Appendix E.

4 METHOD

4.1 SELF-SUPERVISED PRETRAINING

We pretrain BCG-FM with self-supervised contrastive learning on a large corpus of nightly BCG recordings (Figure 1b), following the joint-embedding architecture of Abbaspourazad et al. (2024) as a starting point with minor adaptations to better fit the bed-piezo BCG: a single-channel mechanical signal dominated by respiration on slow timescales and by cardiac motion on fast timescales, recorded ambiently over many nights per user.

Input representation. Each training example is a 60-second BCG window sampled at 64 Hz, decomposed into breathing (using a 0.5 s moving average) and heart (residual) channels that are z-scored per window; unlike band-pass filtering, this split is lossless (the two channels reconstruct the input exactly) and linear-cost for streaming. We did not use synthetic augmentations; instead, positive pairs were constructed from naturally occurring windows from the same user, relying on night-to-night and within-user variability to provide view diversity.

Encoder and projection head. A 1D EfficientNet (Tan and Le, 2019) maps the 2-channel, 3840-sample input to a 256-d feature through a stride-2 stem, 16 MBConv1d blocks with squeeze-and-excitation (ratio 0.25, swish activations), a final 1×1 expansion, and global average pooling. Six stride-2 reductions take the time axis $3840 \to 60$; block widths follow $42 \to 28 \to 35 \to 49 \to 64 \to 78 \to 99 \to 128 \to 256$, kernels alternate between 3 and 5 across stages, and the expansion factor is 7 (1 in the first block). Total: 3.3M parameters. The projection head $g_\theta : \mathbb{R}^{256} \to \mathbb{R}^{128}$ is a 2-layer MLP (hidden 1024, dropout 0.1); downstream tasks use the 256-d pre-projection feature. A momentum teacher (He et al., 2020) (exponential moving average with momentum coefficient m=0.99) provides the target view.

Self-supervised objective and positive-pair construction. InfoNCE with temperature τ=0.04, bidirectional (student↔teacher). KoLeo spread regularizer (Sablayrolles et al., 2019; Oquab et al., 2024) with γ_{koleo}=0.1. Full objective:

$$\mathcal{L} = \tfrac{1}{2}(\mathcal{L}_{1\to2} + \mathcal{L}_{2\to1}) + \gamma_{\text{koleo}}\,\mathcal{L}_{\text{KoLeo}}.$$

Positives are defined at the user-and-session level. Pairing pool of 10^7 rows; anchor at time t_a, target offset $\Delta \sim \mathcal{U}(-90, 90)$ days, $\Delta \neq 0$ (excluding same session-pairs). Unique-user-per-batch constraint. The $\pm$90-day window couples recordings sharing user identity and broad physiological state but differing in nightly context.

Distributed optimization. Training used B=4096 positive pairs across 8 A100-40GB GPUs, with B_{local}=512 per GPU. We used bf16 autocast and computed the loss in fp32. Optimization used Adam with initial learning rate η_0=10^{-3} and weight decay 10^{-4}. The learning rate was first halved at the initial validation plateau, occurring at step 160K, and was subsequently halved every 30K steps. Training ran for 278K steps (88 hours). Validation was performed every 2000 steps using frozen Ridge probes for age, BMI, and identity on a device-disjoint split; age MAE was used for early stopping.

Table 2: Age prediction by downstream head (frozen BCG-FM backbone); test-set metrics.

Downstream head	All ages		Ages 18–70	
	MAE↓	R^2 ↑	MAE↓	R^2 ↑
E2E Supervised	4.411	0.662	4.227	0.662
Ridge (linear probe)	3.803	0.735	3.717	0.716
Transformer + DecadeCE	**3.339**	**0.793**	**3.257**	**0.777**

4.2 Second-stage pretraining with neighbour context

Because participant-level positives treat within-person variation as nuisance, we ask whether new information can be added to BCG-FM while retaining what it encodes. Starting from BCG-FM, a small head on the pooled embedding z_t of minute t predicts the change in the frozen model's embedding between minute t and minutes up to ± 30 minutes away in the same night, $e(t+k) - e(t)$, and a second head regresses the minute's own heart-band log spectrum (16 bands, 0.5–4 Hz). Predicting the *difference* cancels the person component and leaves the state trajectory; a linear head regressing $e(t)$ and an identity head regressing the user's mean embedding retain trait content.

4.3 Downstream evaluation protocol

All downstream evaluations use the *frozen* backbone (Figure 1d): pretraining weights are not updated. Every in-bed minute is mapped to the 256-d pre-projection feature. Evaluation uses user-level metrics throughout. All downstream train, validation, and test splits were user-disjoint, with all sessions from a given user assigned to the same split; age users were SSL-held-out, and disease labels were held out.

Embedding views. The frozen backbone produces one 256-d feature per in-bed minute, aggregated into (i) single minutes for the end-to-end baseline, (ii) full-session minute sequences (mean 663, up to 720 minutes) for the Transformer head, and (iii) user means (minutes $\rightarrow$ session mean $\rightarrow$ user mean) for the Ridge probe.

Downstream heads. To evaluate BCG-FM, we trained two heads on top of the frozen encoder: a **ridge regressor** on user-level embeddings and a session-level **transformer head** for age prediction. The transformer models the continuous overnight BCG sequence using minute-level embeddings with a 2-layer, 4-head architecture (d_{model}=256, FFN dimension 512, dropout 0.2), followed by mean pooling over valid minutes. It was trained with DecadeCE loss (a Gaussian-smoothed cross-entropy over age bins, σ=5) (Zhang et al., 2024), AdamW, weight decay 10^{-4}, batch size 64, cosine learning-rate scheduling, gradient clipping at norm 1.0, and early stopping with patience 8 over at most 60 epochs. Hyperparameters were selected by user-level validation performance. Ridge swept $\alpha \in \{10^{-2}, 10^{-1}, 1, 10, \mathbf{10^2}, 10^3\}$. The transformer swept learning rate $\{10^{-4}, 3{\cdot}10^{-4}\}$, $\sigma \in \{2, 3, 4, \mathbf{5}, 6, 7\}$, layers $\in \{1, \mathbf{2}, 4\}$, and heads $\in \{\mathbf{4}, 8, 16\}$.

4.4 Baselines

We compare BCG-FM frozen embeddings against: (a) **End-to-end supervised**: same encoder trained from scratch on labeled age data (per-minute, user-averaged at inference); (b) **Demographics-only**: age/sex/BMI baseline for diagnosis on internal and external datasets.

5 Experiments and Results

We evaluate BCG-FM along two axes: *what the frozen embeddings encode* (age estimation, disease prediction, label efficiency) and *SSL design choices* (batch-size scaling). Full augmentation ablations are in Appendix A; objective ablations and within-night analyses are in Appendix B.

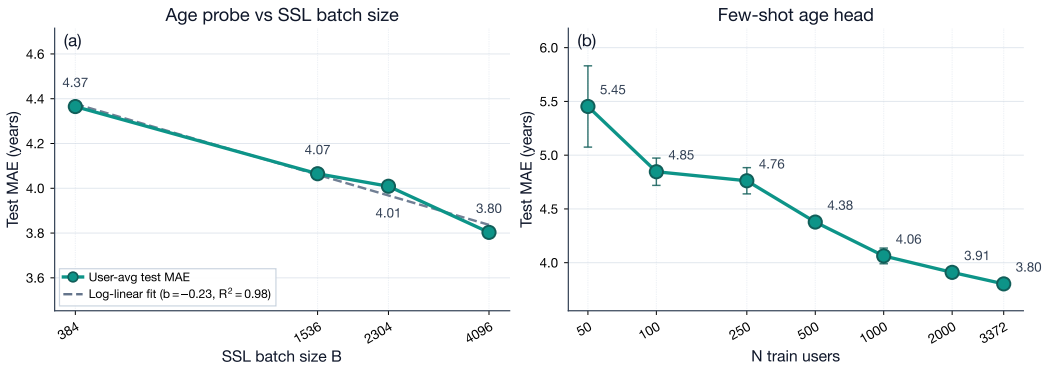

Figure 2: BCG-FM age probe overview: (a) Batch-size scaling — MAE improves log-linearly with batch size (R^2=0.982); (b) Few-shot label efficiency — BCG-FM with 500 users outperforms supervised E2E with 3,372 users.

Table 3: Cross-model comparison of biological age estimation. MAE values are not directly comparable across studies due to differences in cohort composition, age range, and health definitions.

Model	Modality	Setting	Cohort	MAE (yr)	R^2
PpgAge (Miller et al., 2025)	PPG (wrist)	Wearable	Healthy, n=6,728	2.43	0.88
PpgAge (Abbaspourazad et al., 2024)	PPG (wrist)	Wearable	General, n=120,235	3.20	—
ECGFounder (Li et al., 2025)	12-lead ECG	Clinical	Clinical, n=1.8M	8.65	—
ECG-age (Lima et al., 2021)	12-lead ECG	Clinical	Clinical, n=1.55M	8.38	0.71
Brain-age (Brink-Kjær et al., 2022)	PSG (multi-ch)	Clinical	Clinical,n=13k	5.80	0.70
Brain-age (Brink-Kjær et al., 2022)	PSG (ECG)	Clinical	Clinical,n=13k	10.4	—
SleepFM (Thapa et al., 2026)	PSG (multi-ch)	Clinical	Clinical, n=65K	7.33	0.77
BCG-FM (ours, 18–70)	**BCG (bed)**	**Ambient**	**Healthy,** n=4,645	**3.26**	**0.78**

5.1 Age Estimation

BCG-FM's Transformer head achieves **3.26-year MAE** (ages 18–70) and **3.34 MAE** (all ages), with R^2=0.777 and 0.793, respectively (Figure 1e). The Ridge linear probe alone achieves 3.80 MAE (all ages), demonstrating that substantial age-related information is linearly accessible in the 256-d embedding space. The E2E supervised baseline (4.41 MAE) is outperformed by the linear probe, confirming the value of SSL pretraining on BCG data.

5.2 Batch-size scaling

Larger contrastive batches (384 to 4096) improve age, BMI, and identity probes log-linearly; log batch size explains 98% of the variance in age MAE (Figure 2a; Table 7, Appendix B).

5.3 Disease prediction: internal cohort

On the internal cohort, we probe only users who completed the pre-existing condition and sleep condition survey fields. BCG-FM embeddings achieve mean AUROC of 0.684 across 15 self-reported conditions (Table 4, (Figure 1f, left)). Performance is strongest for conditions with established cardiovascular signatures: diabetes (0.853), sleep apnea (0.792), snoring (0.751), heart conditions (0.734), and hypertension (0.810). These conditions also typically have high self-reporting certainty. BCG-FM perform worse on conditions where self-reporting is unreliable or the BCG signal has weaker mechanistic links (thyroid disorder 0.659 or Migraines 0.673). Overall, BCG-FM improves over the demographics-only baseline for 13 of 15 conditions, with a mean AUROC gain of 0.047.

Table 4: Internal disease AUROC. Demo: demographics-only; BCG-FM (w/o): embeddings only; BCG-FM: embeddings + demographics; Δ: BCG-FM gain over Demo.

Disease	Demo	BCG-FM (w/o)	BCG-FM	Δ
Diabetes	0.786	**0.853**	0.852	+0.066
Sleep Apnea	0.751	0.778	**0.792**	+0.041
Snoring	0.706	0.740	**0.751**	+0.046
Heart Conditions	0.669	0.726	**0.734**	+0.065
Hypertension	0.774	0.805	**0.810**	+0.036
Cancer	0.637	**0.683**	0.678	+0.041
Hot Flashes	0.594	0.641	**0.671**	+0.078
Nighttime Hot Flashes	0.596	0.673	**0.676**	+0.080
Migraines	0.642	0.652	**0.673**	+0.032
Asthma/COPD	0.565	0.645	**0.649**	+0.084
Insomnia	0.564	0.617	**0.621**	+0.057
Restless Leg	0.559	0.614	**0.615**	+0.056
Night Sweats	0.547	0.597	**0.600**	+0.053
Irregular Heartbeat	0.474	**0.479**	0.464	−0.010
Thyroid Disorder	**0.694**	0.659	0.680	−0.014
Mean	0.637	0.678	**0.684**	+0.047

Table 5: External disease AUROC. Demo: demographics-only; BCG-FM (w/o): embeddings only; BCG-FM: embeddings + demographics; Δ: BCG-FM gain over Demo.

External Dataset 1 (N=85)					*External Dataset 2 (N=46)*				
Disease	Demo	BCG-FM (w/o)	BCG-FM	Δ	Disease	Demo	BCG-FM (w/o)	BCG-FM	Δ
Abnormal	0.775	0.793	**0.828**	+0.052	AF (user)	0.286	0.569	**0.586**	+0.300
AF	**0.994**	0.941	0.954	−0.040	AF (window)	0.296	0.699	**0.704**	+0.408
HF	0.645	**0.841**	0.822	+0.178					
PACs	0.297	0.285	**0.301**	+0.004					
PVCs	0.526	**0.657**	0.644	+0.119					
Mean	0.647	0.703	**0.710**	+0.062	**Mean**	0.291	0.634	**0.645**	+0.354
Mean (excl. AF)	0.561	0.644	**0.649**	+0.088					

5.4 Disease prediction: external validation

External validation provides evidence that BCG-FM embeddings generalize beyond the internal cohort. On Dataset 1 (N=85), frozen embeddings without demographics achieve AUROC 0.841 for heart failure detection, substantially exceeding the demographics-only baseline (0.645). On Dataset 2 (N=46), AF detection reaches AUROC 0.704 at the window level. Overall, BCG-FM improves over the demographics-only baseline for 6 of 7 external targets (Table 5). The only exception is atrial fibrillation in Dataset 1, where the demographics baseline is already near-perfect (AUROC 0.994), suggesting that this endpoint is dominated by demographic or cohort-structure effects in this dataset. These results are encouraging but must be interpreted cautiously given the small sample sizes; larger multi-site validation studies are needed to establish clinical reliability.

5.5 Label efficiency

A Ridge probe on frozen embeddings reaches 5.45-year age MAE with 50 labeled users and 4.38 with 500 (Figure 2b), outperforming the end-to-end (E2E) baseline trained on all 3,372 (4.41).

5.6 Enriching the embedding with a second pretraining stage

Under the unchanged frozen linear-probe protocol on Twilight (§3.3), the second stage (§4.2) improves every state read-out, most strongly heart rate (−14% MAE), while age is retained (Table 6). Two lighter variants, training details, and full results are given in Appendix C.

Table 6: Frozen linear probes on Twilight before and after the second pretraining stage. Age on validation batches. Single seed.

Read-out	BCG-FM	+ second stage
Sleep staging (4-class) macro AUROC	0.824	**0.831**
Apnea event (per minute) AUROC	0.876	**0.888**
ECG heart rate MAE (bpm) ↓	2.26	**1.94**
Age R^2 / per-person MAE (yr)	0.562 / 6.40	0.559 / 6.43

6 Discussion

6.1 Ambient mechanical biosignals as a foundation model modality

BCG-FM demonstrates that bed-based ballistocardiography can support self-supervised foundation models trained on large-scale longitudinal data, capturing demographic, anthropometric, identity, and disease-related information. Recipes from other biosignals do not transfer directly to BCG, and the pretrained model can be enriched after contrastive pretraining (§5.6). Unlike battery- and adherence-limited wearables, BCG captures the entire sleep period ($\sim$5–10 hours), enabling session-level Transformers across full nights (Table 2: 3.26 vs. 3.72 MAE).

BCG-specific augmentations degrade performance (Appendix A), likely because posture, respiration, sensor coupling, and cardiac mechanics already provide substantial natural variation; synthetic transformations may remove physiologically meaningful signal rather than add diversity. In contrast, larger contrastive batches consistently help: negative diversity is central, and the log-linear trend suggests contrastive BCG learning is predictable and improvable with scale.

BCG-FM compares favorably with reported ECG- and PSG-based age-estimation models (Table 3), including SleepFM (3.26 vs. 7.33 years), which we attribute to four times more pretraining hours and participant-level pretraining across nights; SleepFM's similar R^2 ($\approx$0.77 vs. 0.78) highlights the limits of cross-study comparisons. Likewise, the PPG foundation model of Abbaspourazad et al. (2024) reports diabetes AUROC of 0.789 versus 0.852 for BCG-FM; the comparison is approximate, but a single ambient sensor requiring no user action rivals wearable-grade discrimination for metabolic and cardiovascular conditions.

Finally, datasets with $\geq$30 consecutive nights per participant would let models exploit night-over-night structure informative for aging and disease.

6.2 Limitations

(1) The pretraining corpus comes from a single commercial sensor deployment; generalization across BCG hardware remains to be established at scale. (2) The healthy cohort uses custom health status scores and (3) internal disease labels are self-reported. (4) The external cohorts are small (N=85 and N=46), limiting statistical power and subgroup analyses.

7 Conclusion

We introduced BCG-FM, the first foundation model for ambient mechanical biosignals. Frozen embeddings achieve **3.26-year MAE** for age estimation, the lowest reported for any contactless modality. BCG-FM discriminates diabetes (AUROC 0.852), hypertension (0.810), and sleep apnea (0.792), and generalizes to unseen clinical datasets for abnormal heart rhythms (0.828), atrial fibrillation (0.954 & 0.704), and heart failure (0.822). A label-free second stage further enriches the embeddings with within-night state information ($-$14% HR MAE) while retaining age information. These results establish ambient mechanical biosignals as a viable and clinically relevant modality for health foundation models with potential for low-burden longitudinal health monitoring.

AI USE STATEMENT

We used generative AI tools to polish writing, to draft parts of the manuscript, to retrieve and verify related work and references, and to assist with data-processing code and figure preparation. All AI-assisted output was reviewed and verified by the authors, who take full responsibility for the content of this paper.

ETHICS STATEMENT

All recordings come from a consumer sleep-sensing platform where research use of de-identified data is disclosed and opt-in, covering model training and scientific validation. In dual-occupancy beds each side has its own consented user profile, and no profile is built for anyone else. Users may request deletion at any time and are then excluded from subsequent training; already-trained aggregate weights are not retrained per request. De-identification is not non-identifiability for longitudinal BCG: identity retrieval (rank-1 92.5%) shows stable individual signatures, so we treat raw BCG and embeddings as pseudonymized biometric data and report retrieval as a privacy-risk characterization. Because sensing is passive and contactless, inferred health status could be misused by insurers, employers, or others without consent; we neither deploy nor enable such use, the pretrained encoder is not released, and released data come from a clinical study with explicit consent under a data-use agreement. The user base is skewed, so performance need not transfer to the general or clinical population, and BCG-FM is a research model and screening aid, not a diagnostic device.

REPRODUCIBILITY STATEMENT

Sections 3–4 and Appendices A–F describe the data, architecture, objectives, hyperparameters, and evaluation protocol. The pretraining corpus contains sensitive biometric data and cannot be released; the external datasets are public, and we will release the Twilight dataset with fixed splits and evaluation scripts.

REFERENCES

K. He, et al. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020.

M. Tan and Q. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114, 2019.

S. Abbaspourazad, et al. Large-scale training of foundation models for wearable biosignals. In *ICLR*, 2024.

A. C. Miller, J. Futoma, S. Abbaspourazad, C. Heinze-Deml, S. Emrani, I. Shapiro, and G. Sapiro. A wearable-based aging clock associates with disease and behavior. *Nature Communications*, 2025.

R. Thapa, et al. A multimodal sleep foundation model for disease prediction. *Nature Medicine*, 32:752–762, 2026.

J. Li, et al. An electrocardiogram foundation model built on over 10 million recordings. *NEJM AI*, 2(7), 2025.

E. Coppola, M. Savardi, M. Massussi, M. Adamo, M. Metra, and A. Signoroni. HuBERT-ECG: A self-supervised foundation model for broad and scalable cardiac applications. *medRxiv preprint*, 2024.

A. Pillai, D. Spathis, F. Kawsar, and M. Malekzadeh. PaPaGei: Open foundation models for optical physiological signals. In *ICLR*, 2025.

G. Nie, Q. Zhao, G. Tang, Y. Li, and S. Hong. Artificial intelligence-derived photoplethysmography age as a digital biomarker for cardiovascular health. *Communications Medicine*, 2025.

S. Abbaspourazad, et al. Wearable accelerometer foundation models for health via knowledge distillation. *arXiv preprint arXiv:2412.11276*, 2025.

G. Narayanswamy, et al. Scaling wearable foundation models. In *ICLR*, 2025.

Y. Zhang, et al. SensorLM: Learning the language of wearable sensors. In *NeurIPS*, 2025.

X. Gu, et al. Cardiac health assessment across scenarios and devices using a multimodal foundation model pretrained on data from 1.7 million individuals. *Nature Machine Intelligence*, 8:220–233, 2026.

A. Sablayrolles, M. Douze, C. Schmid, and H. Jégou. Spreading vectors for similarity search. In *ICLR*, 2019.

M. Oquab, et al. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024.

O. T. Inan, et al. Ballistocardiography and seismocardiography: A review of recent advances. *IEEE J. Biomed. Health Inform.*, 19(4):1414–1427, 2015.

C. Brüser, S. Winter, and S. Leonhardt. Robust inter-beat interval estimation in cardiac vibration signals. *Physiological Measurement*, 34(2):123–138, 2013.

R. Yi, M. Enayati, J. M. Keller, M. Popescu, and M. Skubic. Non-invasive in-home sleep stage classification using a ballistocardiography bed sensor. In *Proc. IEEE EMBS Int. Conf. Biomedical & Health Informatics (BHI)*, pages 1–4, 2019.

A. Albukhari, F. Lima, and U. Mescheder. Bed-embedded heart and respiration rates detection by longitudinal ballistocardiography and pattern recognition. *Sensors*, 19(6):1451, 2019.

J. W. Gordon. Certain molar movements of the human body produced by the circulation of the blood. *J. Anat. Physiol.*, 11:533–536, 1877.

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

J. Zhan, Z. Li, X. Wu, et al. A multi-pathology ballistocardiogram dataset for cardiac function monitoring and arrhythmia assessment. *Scientific Data*, 12:963, 2025. DOI: 10.1038/s41597-025-05287-z.

J. Qiu, T. Lyu, L. Liu, et al. A ballistocardiogram dataset with reference ECG signals for bed-based heart rhythm assessment. *Scientific Data*, 12:1651, 2025. DOI: 10.1038/s41597-025-05936-3.

E. M. Lima, et al. Deep neural network-estimated electrocardiographic age as a mortality predictor. *Nature Communications*, 12:5117, 2021. DOI: 10.1038/s41467-021-25351-7.

D. Zhang, Y. She, J. Sun, Y. Cui, X. Yang, X. Zeng, and W. Qin. Brain age estimation from overnight sleep electroencephalography with multi-flow sequence learning. *Nature and Science of Sleep*, 16:879–896, 2024. DOI: 10.2147/NSS.S463495.

G. Nie, X. Fang, G. Tang, Y. Xiao, J. Li, B. Liu, H. Li, and S. Hong. AnyPPG: An ECG-guided PPG foundation model trained on over 100,000 hours of recordings for holistic health profiling. *arXiv preprint arXiv:2511.01747*, 2025.

A BCG-SPECIFIC AUGMENTATION ANALYSIS

We evaluate four domain-plausible augmentations for BCG: clipping, sensor cross-talk, Gaussian noise, and temporal dropout. These augmentations are designed to mimic realistic signal corruption or sensor artifacts. However, unlike prior wearable biosignal models where augmentations often improve invariance, we find that these transformations degrade downstream performance even at low strength.

Clipping. Clipping simulates saturation or amplitude truncation in the BCG waveform.

Sensor cross-talk. Sensor cross-talk simulates leakage or mixing between sensor channels.

Gaussian noise. Gaussian noise simulates random measurement noise.

Temporal dropout. Temporal dropout masks short temporal intervals, simulating missing or corrupted signal segments.

All degrade downstream performance even at low strength (Full 50%: 7.346 MAE; Light 20%: 7.332; Baseline: 7.193; batch size 384). This is in sharp contrast to PPG and ECG foundation models where augmentations consistently improve representation quality.

B Additional objective ablations and within-night analyses

Table 7: Probe metrics vs. SSL batch size (Ridge; RidgeClassifier for binaries; cosine identity retrieval). Binary columns: AUC with pAUC (FPR $< 10\%$) in parentheses.

Batch	Age		Age 18–70		Age ≥ 50	BMI	BMI ≥ 30	Identity	
size	MAE↓	R^2 ↑	MAE↓	R^2 ↑	AUC(pAUC)↑	MAE↓	AUC(pAUC)↑	Rank-1↑	MRR↑
384	4.365	0.665	4.266	0.640	0.928(0.804)	2.303	0.749(0.596)	0.900	0.914
1536	4.065	0.706	3.978	0.685	0.935(0.819)	**2.287**	0.748(0.605)	0.922	0.931
2304	4.009	0.708	3.928	0.684	0.933(0.818)	2.301	0.730(0.615)	0.916	0.927
4096	**3.803**	**0.735**	**3.717**	**0.716**	**0.943(0.832)**	2.299	**0.753(0.618)**	**0.925**	**0.934**

Pretraining objective. In a step-matched comparison at B=384, we augmented the contrastive student with an auxiliary masked-reconstruction head: 8 input patches (480 samples each) are replaced with a learned mask token and a head predicts their log-power spectrum (λ=0.1; teacher unmasked). Validation age-probe MAE was 7.133 yr for contrastive only, 7.144 yr with the auxiliary head at 0% masking (performance-neutral), and 8.643 yr at 75% masking (+1.51 yr). A standalone JEPA-style ViT, trained without embedding collapse, reached only 9.44 yr. Standard PPG/ECG augmentations degrade performance monotonically with dose (7.193→7.346 yr; Appendix A). These are comparisons against contrastive learning *plus* reconstruction; a matched standalone masked-reconstruction baseline remains future work.

Co-sleeper hard negatives. Ambient sensing offers negatives unavailable to wearables: a co-sleeping partner recorded simultaneously by the same bed system. Filling $\approx$35% of each batch with co-sleeper pairs, without reducing participant diversity, improved validation age MAE by 0.05 yr. Together with log-linear batch-size scaling (Section 5.2), this suggests that both the quantity and the hardness of negatives drive BCG representation quality.

Within-night physiology. To test whether participant-level positives suppress within-person dynamics, we probe frozen minute-level BCG-FM embeddings on the internal Twilight dataset (§3.3). With subject- and bed-disjoint 5-fold linear probes, BCG-FM reaches a 3-class (wake/NREM/REM) macro AUROC of 0.867, a 4-class (wake/light/deep/REM) macro AUROC of 0.824 (0.841 with $\pm$2/$\pm$5-minute context means), a 3-class AUROC of 0.884 with $\pm$5-minute context and 0.894 with a heavier night-level Transformer read-out with Viterbi smoothing (not comparable to the linear probes in Table 6), a per-minute apnea-event AUROC of 0.876, and an ECG heart-rate MAE of 2.26 bpm from a ridge probe. Same-night minute-level cosine similarity is moderate (mean 0.452) and highly variable (0.00–0.95; 10 participants, 61 sessions), whereas session-level identity separation is strong, indicating that stable traits concentrate in session-level aggregates while minute-level embeddings retain dynamic physiology. Section 5.6 shows that this state information can be increased further.

Night-to-night variation. Across 368,171 sessions from 68,009 participants, predicted age shows a weekly cycle of 0.26 yr peak-to-trough: work-week nights read older and weekend nights younger (Mon +0.10, Tue +0.11, Wed +0.04, Thu +0.01, Fri −0.11, Sat −0.15, Sun −0.05 yr relative to the mean), indicating sensitivity to behavioural night-to-night changes.

C NEIGHBOUR-CONTEXT SECOND-STAGE PRETRAINING: DETAILS

Objective. Let f be the student encoder initialised from BCG-FM, $z_t = f(x_t)$ the pooled 256-d embedding of minute t, and $e(\cdot)$ the frozen BCG-FM embedding, precomputed for 227M pretraining minutes. The loss is

$$\mathcal{L} = \mathcal{L}_{\text{nb}} + 0.5\,\mathcal{L}_{\text{pin}} + 0.5\,\mathcal{L}_{\text{id}} + 0.1\,\mathcal{L}_{\text{pv}}\ [+\,0.5\,\mathcal{L}_{\text{spec}}],$$

where $\mathcal{L}_{\text{nb}}$ is the MSE between $g_k(z_t)$ (MLP $256{\to}512{\to}6{\times}256$) and the per-dimension standardised difference $e(t{+}k) - e(t)$ for $k \in \{\pm1, \pm2, \pm5\}$ min (NC-5, NC-5+HS) or $k \in \{\pm1, \pm2, \pm5, \pm10, \pm20, \pm30\}$ min (NC-30+HS, the configuration in Table 6), masking neighbours that are missing or outside the night; $\mathcal{L}_{\text{pin}}$ is $(1 - \cos)+$ normalised MSE between a linear map Wz_t and $e(t)$ on the live dimensions; $\mathcal{L}_{\text{id}}$ has the same form for an identity head regressing the user's mean embedding; $\mathcal{L}_{\text{pv}}$ is a pool-variance floor; and the $\mathcal{L}_{\text{spec}}$ (NC-5+HS and NC-30+HS only) regresses the minute's own 16-band heart-band (0.5–4 Hz) log spectrum. Predicting the neighbour's raw embedding instead of the difference carried no new signal in a pilot and degraded heart rate (2.26→2.9 bpm).

Motivation. Participant-level contrastive pretraining rewards what is stable within a person and treats within-night variation as nuisance, so we hypothesized that the embedding retains unused capacity for physiological state. The second stage tests whether this capacity can be filled with new, label-free information while the existing trait content is preserved by the retention heads.

Training. Initialisation from BCG-FM; same inputs (breathing and heart channels, 60-s windows at 64 Hz, no masking); 8×A10G GPUs; global batch 1024; AdamW, learning rate $2.5{\times}10^{-4}$, weight decay 10^{-4}, halved at 40k steps; trunk frozen for 2k steps then ramped in over 4k; neighbour loss warmed up over 5k. The checkpoint rule was fixed before any evaluation: the latest checkpoint before any training-loss spike whose validation age R^2 stays ≥ 0.557 (BCG-FM -0.005), giving steps 52k (NC-5), 32k (NC-5+HS), and 42k (NC-30+HS).

Table 8: Full results for the three second-stage variants (single seed each). NC-5: neighbour offsets up to ±5 min; NC-5+HS: NC-5 plus heart-band spectrum (HS) head; NC-30+HS: offsets up to ±30 min plus HS head.

Frozen linear read-out	BCG-FM	NC-5 @52k	NC-5+HS @32k	NC-30+HS @42k
4-class macro AUROC	0.824	**0.833**	0.830	0.831
4-class macro AUROC + context	0.841	**0.850**	0.847	0.847
3-class macro AUROC	0.867	**0.873**	0.871	0.872
Apnea AUROC (probe / + context)	0.876 / 0.881	**0.891 / 0.899**	0.886 / 0.895	0.888 / 0.895
ECG heart rate MAE (bpm) ↓	2.26	2.11	1.95	**1.94**
Age R^2 / per-person MAE (yr)	**0.5623 / 6.40**	0.5605 / 6.42	0.5582 / 6.44	0.5590 / 6.43
Sex accuracy	0.685	0.685	0.684	0.684

D THE TWILIGHT DATASET

Collection. Twilight contains 195 in-lab nights from 195 participants, recorded between December 2025 and March 2026 on 10 lab beds fitted with the same generation of bed sensor. Each night has bed BCG recorded simultaneously with full polysomnography. Sensors are three (92 nights) or four (103 nights) distributed piezoelectric elements per bed side, sampled at 100 Hz together with a combined piezo channel; BCG-FM reads the combined channel, resampled to 64 Hz in 60-second windows, exactly as in pretraining. 167 nights were recorded with a single sleeper and 28 with a bed partner present, with PSG on one side, providing realistic dual-occupancy conditions. No recording from these 10 beds is part of the pretraining data.

Recordings and labels. The dataset comprises 1,944 recorded hours (mean 7.9 h per night in bed; mean total sleep time 6.2 h, SD 1.2, range 1.7–9.0), of which 1,545 hours carry PSG sleep-stage labels at 1-minute resolution (92,673 minutes). The stage distribution over scored minutes is N2 41%, wake 22%, N1 14%, REM 12%, and N3 11%; we evaluate 4 classes (wake / light = N1+N2 /

deep = N3 / REM) and 3 classes (wake / NREM / REM). Respiratory events (duration $\geq$10 s, median 19 s) from PSG define a per-minute apnea label (a minute is positive if any event overlaps it; 26.6% prevalence). The cohort spans the full range of sleep-disordered breathing: median apnea–hypopnea index 15.9 events/h (IQR 4.4–38.5), with 54 participants below 5, 42 between 5 and 15, 36 between 15 and 30, and 63 at 30 or above. Per-minute reference heart rate is derived from R-peaks in the PSG ECG (100 Hz); minutes between 30 and 140 bpm with an ECG quality index $\geq$0.9 are used.

Benchmark protocol. Probes use five fixed folds that are disjoint by both subject and lab bed. Tasks are 4- and 3-class sleep staging (macro AUROC), per-minute apnea-event detection (AUROC), and ECG heart-rate regression (MAE); read-outs are logistic or ridge regression on frozen minute embeddings, optionally with $\pm$2/$\pm$5-minute context means. Paired differences between models use a 2000-resample subject bootstrap. Baseline results for BCG-FM and the second-stage variants are in Table 8.

Release. We will release raw BCG and PSG sleep-stage labels for all 195 nights with age bands, organized into fixed training (113 nights) and test (82 nights) splits, together with evaluation scripts and linear-probe baselines for sleep staging, under controlled access.

E Health Status Score (HSS) definition

The Health Status Score is a three-axis composite used to define the healthy reference cohort for age estimation. It integrates self-reported survey data with objectively measured sleep architecture and exercise metrics.

E.1 Axis 1: Survey HSS (0–100 points)

The survey HSS comprises four components scored from a one-time health questionnaire:

Component	Range	Scoring
Condition burden	0–40	40 = no conditions; 30 = minor only (migraines, allergies); 20 = moderate (anxiety, depression); 5 = major (diabetes, heart disease, sleep apnea); 15 = unanswered (capped at Tier 3)
Sleep quality	0–20	Self-reported satisfaction, onset latency, nighttime awakenings
Lifestyle	0–20	Exercise frequency, alcohol, smoking, caffeine
BMI	0–20	20 = normal (18.5–25); 14 = overweight (25–30); 6 = obese I (30–35); 0 = obese II+ ($>$35); 8 = missing

Approximately 46.9% of survey respondents left the pre-existing conditions field unanswered. These users receive a default score of 15/40 on that component and are capped at Tier 3 regardless of other scores, as their health status is genuinely unknown rather than intermediate.

E.2 Axis 2: Sleep architecture adjustment (-15 to $+8$ points)

For users with device-measured sleep data, an objective sleep adjustment modifies the survey HSS based on six metrics: sleep efficiency, onset latency, deep sleep percentage, REM percentage, wake-after-sleep-onset, and total sleep time. Thresholds are age-adjusted. The asymmetric range (larger penalty than bonus) reflects the design decision that objectively poor sleep should override self-reported health, while good sleep provides only modest confirmation.

E.3 Axis 3: Exercise adjustment (-5 to $+10$ points)

For the subset of users with linked activity data ($\sim$60%), objective workout metrics override self-reported exercise frequency. Users meeting WHO physical activity guidelines ($\geq$150 min/week

moderate or $\geq$75 min/week vigorous) receive positive adjustments; sedentary users despite claiming exercise receive negative adjustments.

E.4 Tier assignment

Final tier assignment uses a 2D matrix of HSS grade and data confidence level:

HSS Grade	Gold	Silver	Bronze	Insufficient
A (80–100)	Tier 1	Tier 1	Tier 2	Tier 3
B (60–79)	Tier 1	Tier 2	Tier 2	Tier 3
C (40–59)	Tier 2	Tier 3	Tier 3	Exclude
D (20–39)	Tier 3	Tier 4	Tier 4	Exclude
F (0–19)	Tier 4	Tier 4	Exclude	Exclude

Data confidence reflects the number and recency of objective data streams available: Gold requires both sleep and exercise data within 90 days of the survey; Silver requires sleep data only; Bronze has survey only; Insufficient lacks valid survey completion. At production scale ($\sim$160K scored users), this yields an estimated 13–20% Tier 1 population (21,000–32,000 users), which is 3–5$\times$ the PpgAge reference cohort of 6,728.

Only Tier 1 users enter the healthy age-estimation cohort. The 9,151 users used in this study represent the intersection of Tier 1 assignment with sufficient BCG recording sessions ($\geq$3 nights).

F Additional dataset details

Existing assets and licenses. External Dataset 1 (Zhan et al., 2025) and External Dataset 2 (Qiu et al., 2025) are publicly available Figshare datasets released under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. We cite the original dataset descriptors and use the datasets in accordance with their stated terms. No third-party pretrained model weights were used.

Table 9: Health Status Score (HSS) tier distribution. Tier 1 is used for age estimation; Tier 2–4 and excluded users are available for disease prediction.

Tier	Users	% of survey	Description
1 (Healthy reference)	∼21–32K	13–20%	Grade A/B + Gold/Silver confidence
2 (Mostly healthy)	∼32–40K	20–25%	High survey score, lower data confidence
3 (Mixed/unknown)	∼40–56K	25–35%	Moderate score or unanswered conditions
4 (Clearly unhealthy)	∼16–24K	10–15%	Low survey score
Excluded	∼24–32K	15–20%	Insufficient data or extreme BMI

Table 10: Validation **age-probe** metrics for different levels of augmentation probability (batch size 384). Val MAE in years; lower MAE and higher R^2 are better. The augmentation SSL run is trained up to 146 000 steps and comparisons were done here for the baseline model too.

Augmentation fraction	Val MAE↓	Val R^2 ↑
Full 50%	7.346	0.449
Light 20%	7.332	0.448
Baseline (no robustness aug.)	7.193	0.467