Time-varying Representations of Longitudinal Biosignals using Self-supervised Learning

Sam Perochon[*]
Centre Borelli
Ecole Normale Supérieure Paris-Saclay
sam.perochon@ens-paris-saclay.fr

Salar Abbaspourazad
Apple
salarabb@apple.com

Joseph Futoma
Apple
jfutoma@apple.com

Guillermo Sapiro
Apple
gsapiro@apple.com

Andrew C. Miller
Apple
acmiller@apple.com

Abstract

Many chronic diseases exhibit complex and slow time courses, and in asymptomatic stages it may be possible to detect signs of disease through longitudinal monitoring with wearables. Properly accounting for temporal dependencies in the learned representations of wearable biosignals is crucial to better characterize the progression of disease and improve human health. While previous research has demonstrated that informative representations of wearables-derived biosignals offer much promise in various medical applications, the limited longitudinal scale of most existing wearables datasets has hindered the development of computational and evaluation frameworks that capture these temporal variations with appropriately fine granularity. To address this, we examine the implicit integration of biosignal timestamps in contrastive self-supervised learning when defining the positive pairs of joint-embedding architectures, enforcing physiological consistency by encouraging positive pairs to be close in time. We demonstrate that using this temporal knowledge during pre-training leads to representations more sensitive to time, as they are better able to predict the time of day and overnight binary sleep-wake stages. We also show that these time-aware representations can improve biomarker monitoring, applying them to predict changes in cardiopulmonary fitness, diabetes status, body mass index, and cardiovascular risk. Crucially, we emphasize the importance of a longitudinal within-subject evaluation rather than the more common cross-sectional across-subject evaluation. Our results suggest that time-varying representations can improve the accuracy of health monitoring using wearable-based biosignals, and open the door for future applications of more time-aware representation learning.

1 Introduction

Cardiovascular diseases are the leading cause of death worldwide, representing 32.8% of all-cause deaths in 2019, and are projected to affect around 40.5% of the US population by 2030 [1, 2]. Wearable biosignals such as the photoplethysmogram (PPG) offer the potential for earlier screening and continuous monitoring of cardiovascular, respiratory, and metabolic conditions [3, 4, 5]. PPG measures the volumetric variations of blood circulation in the microvasculature of the wrist by emitting light at the surface of the skin [6]. Biosignals such as PPG may vary over time due to circadian

[*]Work done while at Apple.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

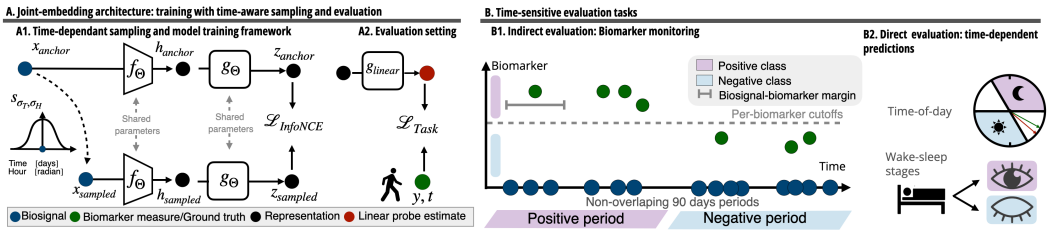

Figure 1: **A) Illustration of the proposed time-aware sampling for joint-embedding architecture, and (B) tasks used to assess biosignals representations time-sensitivity.** Motivations and proposed solutions to the time-awareness desiderata are illustrated in Figure 2 of Appendix A.

rhythms and the complex dynamics of health trajectories, inducing important intra-subject variability [7]. The morphological changes in PPG waveforms can stem from diverse factors, including the aging process, influence of stressors, changes in physical fitness, onset of diseases, natural biological periodicity, and therapeutic interventions [8, 3, 7, 9, 10]. For instance, blood pressure, an important cardiovascular risk factor, exhibits circadian rhythms [11].

Concurrently, recent advances in self-supervised representation learning have enabled the encoding of multivariate physiological time series into rich and compact representations without requiring labels, lessening the need for expensive manual labeling by trained human experts [12, 13]. However, collecting large and *longitudinal* biomedical data is slow and costly, and most work has focused on the evaluation of biosignals representations in cross-sectional settings with static outcomes [4, 12], or with limited sample sizes in controlled longitudinal settings [14].In this work, we investigate implicit use of timestamps when encoding biosignal with a joint-embedding architecture in contrastive self-supervised learning. We focus specifically on evaluating extensively four distinct methods for sampling positive pairs. To assess the time-awareness of learned representations, we introduce a novel evaluation framework, leveraging wearables-derived PPG data from the large longitudinal Apple Heart and Movement Study (AHMS) [15]. Related work of this research can be found in Appendix B. In this work, we investigate implicit use of timestamps when encoding biosignal with a joint-embedding architecture in contrastive self-supervised learning. We focus specifically on evaluating extensively four distinct methods for sampling positive pairs. To assess the time-awareness of learned representations, we introduce a novel evaluation framework, leveraging wearables-derived PPG data from the large longitudinal Apple Heart and Movement Study (AHMS) [15]. Related work of this research can be found in Appendix B.

Our main contributions are as follows: (1) We develop new time-dependent sampling strategies of positive pairs in self-supervised learning (SSL) for biosignals. (2) We show that our time-dependent sampling results in learned representations that capture information about the time of day and circadian effects, as they can better predict the hour of the day as well as binary sleep-wake stages (using wearables-derived estimates). (3) We demonstrate the benefits of our approach by evaluating our ability to use only PPG to predict various health metrics over time: diabetes status (using hemoglobin A1C), cardiopulmonary fitness (using wearables-derived submaximal VO2max estimates), body mass index (BMI), and cardiovascular risk (as estimated by the Framingham risk score). (4) We highlight the importance of evaluating biomarker monitoring in a longitudinal or within-subject manner, compared to a more standard cross-sectional or across-subject manner. We show that standard cross-sectional evaluation suggests that our proposed time-aware positive pair sampling does not improve much over a uniform sampling baseline, while a within-subject evaluation shows modest improvements. To the best of our knowledge, this is the first study to use prediction of the hour of day from biosignal representations as a proxy to quantify sensitivity to circadian rhythms, and to predict cardiovascular risk directly from PPG.

2 Time-dependent positive pairs sampling for joint embedding architecture

Problem definition. For a participant, we define an observable physiological time series as $t :\rightarrow \boldsymbol{x}(t)$, with $t \in [0, T]$, and $\boldsymbol{X}_t = (\boldsymbol{x_1}, ..., \boldsymbol{x_n})$, with $\boldsymbol{x_i} \in \mathbb{R}^{C,N}$, the physiological multivariate time series sampled at irregular timestamps $(t_1, ..., t_n)$, with time-shifted distributions indexed by t. The number of channels C depends on the biosignal and the fixed length of the time series N depend

on the sensor sampling frequency and duration of measurement. The timestamp t_i of a sample $\boldsymbol{x_i}$ is described by the tuple (d_i, θ_i), with d_i the number of days passed since the first measurement at $d_1 = 0$, and $\theta_i \in [0, 24)$ encoding the hour of day. Both time scales are relevant for biological signals. We define a time-dependent generic outcome trajectory as $t \to y(t)$, and denote as $\boldsymbol{Y}_t = (y_1, ..., y_M)$ the available measurements, assumed continuous and univariate in this work. In practice, the number of available measurements M and the enrollment duration T vary across participants.

The deep learning SSL encoder is denoted as f_Θ, with parameters Θ, and maps the multivariate time series $\boldsymbol{x_i}$ onto a lower-dimensional representation, $\boldsymbol{h_i} \in \mathbb{R}^D$, with $D << CN$.

Deep neural network encoder. We used a joint-embedding architecture for the training of f_Θ, similar to the one presented in [12]. The backbone encoder was an EfficientNet-style 1D convolutional neural network with 16 mobile-inverted bottleneck blocks with squeeze-and-excitation [16]. We used the same augmentation module $A(\cdot)$, applying a stochastic sequence of time-series augmentations, including time and magnitude warp, channel swap, crop, and Gaussian noise [17]. We trained the model via gradient-descent using the InfoNCE contrastive loss to prevent representation collapse [18] [19]. We used a projection head during training after the backbone encoder, to compute the loss, as it was shown to improve representation quality [18]. Similar to prior work in [12], we used Kozachenko-Leonenko (KoLeo) regularization to promote variations in different features of the representations. The model architecture and sampling workflow are presented in Figure 1-A, and additional implementation details in Appendix C.1 with full backbone architecture in Appendix C.1.

Sampling of positive pairs. Given an anchor segment $\boldsymbol{x_i}$ recorded at time (d_i, θ_i), we resample a new positive pair for $\boldsymbol{x_i}$ each time we see it during training, using a sampling distribution $s(\cdot)$ defined over the set of remaining samples $\boldsymbol{X}_t \setminus \{x_i\}$ from the same participant. We compare the standard time-independent uniform sampling, that is, $s = s_{uniform} \hookrightarrow \mathcal{U}$, with a sampling dependent on (i) time (days), (ii) hour of day, or (iii) both of these. These are motivated by the assumptions that $d_i \approx d_j \Rightarrow y(d_i) \approx y(d_j)$ and/or $\theta_i \approx \theta_j \Rightarrow y(\theta_i) \approx y(\theta_j)$. Thus, we attempt to guide the model with an inductive bias that biosignals from similar points in time and/or time of day should be more similar than biosignals from the same subject but sampled at a different time of day or very far apart in time. We use Gaussian distributions parameterized by a mean value dependent on the anchor timestamp, and a constant standard deviation value (tuned per downstream task). See additional implementation details in Appendix C.2, and an illustration of the sampling mechanism in Figure 4.

3 Experiments

Experiments were performed using PPG collected via Apple Watch from the Apple Heart and Movement Study (AHMS) [15] (see more details in Appendix D).

Training dataset. We curated 9.99M PPG segments from 172,318 participants for training of the representations. PPG segments were drawn from the full dataset of all PPGs in AHMS based on the following criteria: 1) each participant contributed at least four segments to the pre-training dataset, and 2) the number of segments per participant was as uniform as possible.

Direct evaluation of the representations time-sensitivity: To assess how well different representations are able to encode information about the time of day and other circadian patterns, we created a time-of-day dataset and a binary wake-sleep dataset, where we sampled up to 25 segments per participant. Our evaluation targets are whether a given PPG was taken during the day (10AM-6PM) or night (10PM-6AM), and whether a given PPG during an overnight sleep period was from a sleep or wake period.

Evaluation of the time-sensitivity using biomarker monitoring performances: To assess how well our representations track changes in biomarkers (which may depend on both time-of-day and absolute time), we created four longitudinal datasets to track (i) cardiovascular risk (Framingham) score, (ii) VO2Max, (iii) hemoglobin A1C (from clinical health records), and (iv) BMI (self-reported, or from smart scales). We dichotomized the outcome values using diagnosis status cutoffs provided by published best practice medical guidelines for each outcome, see Figure 5 in Appendix E. Our base units of observation for evaluation are non-overlapping 90-day periods of time per subject, where each 90-day period is labeled according to the most common class label per outcome. For each outcome, we separate the range of values into two categories to create binary labels: we create a "high-contrast" (HC) version by excluding the *pre-diabetes*, *borderline*, *low-medium* and *medium-high*, and *overweight* categories in the A1C, Framingham, VO2Max, and BMI scales, respectively, in order

to create higher-signal tasks. We also report the “low-contrast" (LC) versions of each task, which are binary tasks that make use of all bins for each outcome.

For both evaluations, we used linear probing with ridge-penalized linear models on the representations produced by the trained biosignal encoders (without the projection head) [20]. For each downstream task dataset, we defined a different 80/20% train/test split, with splits defined on a per-subject basis. We quantify performance with the area under the ROC curve (AUC). To quantify uncertainty, we bootstrap resampled each test set 200 times, and below each method report the standard error of the mean. Methods that are non-inferior to the best method (based on a paired one-sided Wilcoxon signed-rank test, with $p > .05$) are underlined. See illustrations of the tasks Figure 1-B and additional information in Appendix E.

4 Lessons learned

We compare the performances of the three proposed time-dependent sampling strategies are with the baseline time-independent uniform sampling. All within-subject results are in Table 1, while across-subject results are in Table 3 in Appendix F.

Time-dependent sampling of positive pairs improved sensitivity to daily PPG variations. Performance for predicting day vs night improved when encouraging representations drawn around the same time of day ($\sigma_H > 0$) to be similar. Unsurprisingly, performance did not improve if positive samples were only drawn close in time but not in hour of the day (i.e. $\sigma_H \to +\infty$ and $\sigma_T > 0$ finite). We emphasize that time-of-day prediction is only a proxy task to quantify and sanity check whether circadian information is retained in the representations.

Similarly, all time-dependent sampling strategies outperformed time-independent sampling for the wake-sleep classification, with the hour-based sampling being the best. Note that the “wake" classes are brief awake stages during overnight sleep, which makes this task intrinsically more difficult than the simpler time of day prediction. However, the motivation for this task is similar – the wake-sleep task itself is not necessarily clinically impactful, but helps confirm that the representations are able to track changes within the day.

Time-dependent sampling of the positive pairs improved the sensitivity to change of biomarkers status, when evaluating within-subject. When evaluating within-subject, for every task there is at least one novel time-dependent sampling strategy has statistically significant improvement over the baseline uniform sampling, although the gains are sometimes small (Table 1). Generally for these within-subject evaluations, the combination of time- and hour-based sampling (TH) performs best, as across the 8 biomarker tasks (4 outcomes, each with a low and high contrast version), it is best in 5 cases. Although no single sampling strategy dominates, emphasizing the need for careful tuning for the task at hand, it seems that the combination sampling strategy is a good default choice. We see the strongest performance gains for our novel time-dependent sampling strategies on the A1C and Framingham tasks, although the high-contrast BMI and VO2max tasks also show modest gains. The improvements yield up to a 5% absolute increase in AUC over the baseline, a gain that holds practical significance for real-world applications.

We emphasize that while we see improvements in our within-subject evaluations, the strong across-subject evaluations of the baseline (Table 3) are maintained, or slightly improved in some cases, with our time-dependent sampling strategies. The across-subject performance of our models are significantly better than other reported numbers in the literature for representation learning on sensor signals [21, 22, 23].

This is an important takeaway from our results: since our goal was to develop better methods for representation learning that track changes over time, we must evaluate these methods within-subject rather than across-subject to observe these improvements. It is interesting to note the drastic absolute differences in magnitude between the within-subject results vs the across-subject results. While PPG can very well discriminate between subjects with differing levels of metabolic health, fitness, and cardiovascular risk, it is much more challenging to distinguish between changes in these types of outcomes within a single individual. Limitations and directions of improvement are presented in Appendix G. We believe that future work can continue to improve on the ability to create useful time-aware representations for biosignals, unlocking new and impactful downstream clinical applications.

Dataset	Task	Positive pairs sampling			
		U	T	H	TH
Time-of-day		0.810 (.000)	0.807 (.000)	0.827 (.000)	**0.829** (.000)
Sleep-Wake		0.833 (.000)	0.854 (.000)	**0.862** (.000)	0.861 (.000)
A1C	LC	0.646 (.003)	**0.664** (.003)	0.649 (.003)	0.646 (.003)
	HC	0.676 (.006)	0.676 (.006)	**0.723** (.006)	0.709 (.006)
BMI	LC	0.757 (.001)	0.759 (.001)	0.752 (.001)	**0.760** (.001)
	HC	0.805 (.003)	0.840 (.003)	0.835 (.003)	**0.851** (.002)
VO2Max	LC	0.634 (.000)	0.638 (.001)	0.637 (.000)	**0.639** (.000)
	HC	0.721 (.002)	**0.733** (.002)	0.713 (.002)	0.730 (.002)
Framingham	LC	0.609 (.003)	0.635 (.003)	0.631 (.003)	**0.667** (.003)
	HC	0.646 (.005)	0.657 (.005)	0.641 (.005)	**0.674** (.005)

Table 1: **Within-subject performance (AUROCs) of PPG representations using proposed time-aware sampling of positive pairs.** Best method is in bold, bootstrap standard errors are in parenthesis, and underlined methods are statistically non-inferior to the best method. In all cases, time-aware representations were significantly better than the baseline.

Acknowledgments and Disclosure of Funding

We would like to thank participants in Apple Heart and Movement Study, Calum MacRae, MD, PhD, and study staff at The Brigham and Women's Hospital, a Harvard affiliate, without whom this work would not have been possible.

The Apple Heart and Movement study was conducted in collaboration between Apple and Brigham and Women's Hospital. Study participants were all Apple Watch users at least 18 years old residing in the United States, and provided informed consent electronically in the Apple Research app. The study was approved by the Advarra Central Institutional Review Board, and registered to ClinicalTrials.gov (Identifier: NCT04198194) [15].

We have no external funding or competing interests for this work to declare.

References

[1] Paul A Heidenreich, Justin G Trogdon, Olga A Khavjou, Javed Butler, Kathleen Dracup, Michael D Ezekowitz, Eric Andrew Finkelstein, Yuling Hong, S Claiborne Johnston, Amit Khera, Donald M Lloyd-Jones, Sue A Nelson, Graham Nichol, Diane Orenstein, Peter W F Wilson, Y Joseph Woo, American Heart Association Advocacy Coordinating Committee, Stroke Council, Council on Cardiovascular Radiology and Intervention, Council on Clinical Cardiology, Council on Epidemiology and Prevention, Council on Arteriosclerosis, Thrombosis and Vascular Biology, Council on Cardiopulmonary, Critical Care, Perioperative and Resuscitation, Council on Cardiovascular Nursing, Council on the Kidney in Cardiovascular Disease, and Council on Cardiovascular Surgery and Anesthesia, and Interdisciplinary Council on Quality of Care and Outcomes Research. Forecasting the future of cardiovascular disease in the united states: a policy statement from the american heart association. *Circulation*, 123(8):933–944, March 2011.

[2] IHME. Global burden of disease study 2019 (gbd 2019). *Global Burden of Disease Collaborative Network.*, 2020.

[3] Denisse Castaneda, Aibhlin Esparza, Mohammad Ghamari, Cinna Soltanpur, and Homer Nazeran. A review on wearable photoplethysmography sensors and their potential future applications in health care. *Int. J. Biosens. Bioelectron.*, 4(4):195–202, August 2018.

[4] Jiewei Lai, Huixin Tan, Jinliang Wang, Lei Ji, Jun Guo, Baoshi Han, Yajun Shi, Qianjin Feng, and Wei Yang. Practical intelligent diagnostic algorithm for wearable 12-lead ECG via self-supervised learning on large-scale dataset. *Nat. Commun.*, 14(1):3741, June 2023.

[5] John Allen, Klaus Overbeck, Gerard Stansby, and Alan Murray. Photoplethysmography assessments in cardiovascular disease. *Measurement and Control*, 39(3):80–83, 2006.

[6] Toshiyo Tamura, Yuka Maeda, Masaki Sekine, and Masaki Yoshida. Wearable photoplethysmographic sensors–past and present. *Electronics*, 3(2):282–302, 2014.

[7] Xiang Lan, Dianwen Ng, Shenda Hong, and Mengling Feng. Intra-inter subject self-supervised learning for multivariate cardiac signals. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(4):4532–4540, 2022.

[8] Ian Shapiro, Jeff Stein, Calum MacRae, and Michael O'Reilly. Pulse oximetry values from 33,080 participants in the apple heart & movement study. *npj Digital Medicine*, 6(1):134, July 2023.

[9] Luc Berthouze, Leon M James, and Simon F Farmer. Human eeg shows long-range temporal correlations of oscillation amplitude in theta, alpha and beta bands across a wide age range. *Clin. Neurophysiol.*, 121(8):1187–1197, August 2010.

[10] Dimitris Spathis, Ignacio Perez-Pozuelo, Soren Brage, Nicholas J. Wareham, and Cecilia Mascolo. Self-supervised transfer learning of physiological representations from free-living wearable data. In *Proceedings of the Conference on Health, Inference, and Learning*, page 69–78, New York, NY, USA, 2021. CHIL '21, Association for Computing Machinery.

[11] Lauren G Douma and Michelle L Gumz. Circadian clock-mediated regulation of blood pressure. *Free Radic. Biol. Med.*, 119:108–114, May 2018.

[12] Salar Abbaspourazad, Oussama Elachqar, Andrew C. Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. Large-scale training of foundation models for wearable biosignals. In *International Conference on Learning Representations (ICLR; IN REVIEW)*, 2023.

[13] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[14] Zheng Xu, Nicole Zahradka, Seyvonne Ip, Amir Koneshloo, Ryan T Roemmich, Sameep Sehgal, Kristin B Highland, and Peter C Searson. Evaluation of physical health status beyond daily step count using a wearable activity sensor. *npj Digital Medicine*, 5(1):164, 2022.

[15] Apple. Apple heart & movement study. ClinicalTrials.gov Identifier: NCT04198194, 2019.

[16] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[17] Brian Kenji Iwana and Seiichi Uchida. An empirical survey of data augmentation for time series classification with neural networks. *PLoS One*, 16(7):e0254841, July 2021.

[18] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020.

[19] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[20] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

[21] Dimitris Spathis, Ignacio Perez-Pozuelo, Soren Brage, Nicholas J. Wareham, and Cecilia Mascolo. Self-supervised transfer learning of physiological representations from free-living wearable data. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 69–78, April 2021. arXiv:2011.12121 [cs, eess].

[22] Xian Wu, Chao Huang, Pablo Roblesgranda, and Nitesh Chawla. Representation Learning on Variable Length and Incomplete Wearable-Sensory Time Series, May 2020. arXiv:2002.03595 [cs, eess, stat].

[23] Haraldur T. Hallgrímsson, Filip Jankovic, Tim Althoff, and Luca Foschini. Learning Individualized Cardiovascular Responses from Large-scale Wearable Sensors Data, December 2018. arXiv:1812.01696 [cs, stat].

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[25] OpenAI. GPT-4 Technical Report, March 2023.

[26] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[27] 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 *International Conference on Machine Learning*, pages 1298–1312. PMLR, 2022.

[28] Han Liu, Zhenbo Zhao, and Qiang She. Self-supervised ECG pre-training. *Biomedical Signal Processing and Control*, 70:103010, 2021.

[29] Temesgen Mehari and Nils Strodthoff. Self-supervised representation learning from 12-lead ECG data. *Computers in Biology and Medicine*, 141:105114, 2022.

[30] Hubert Banville, Omar Chehab, Aapo Hyvärinen, Denis-Alexander Engemann, and Alexandre Gramfort. Uncovering the structure of clinical eeg signals with self-supervised learning. *J. Neural Eng.*, 18(4):046020, March 2021.

[31] Ramin Ghorbani, Marcel J.T. Reinders, and David M.J. Tax. Self-supervised ppg representation learning shows high inter-subject variability. In *Proceedings of the 2023 8th International Conference on Machine Learning Technologies*, Icmlt '23, page 127–132, New York, NY, USA, 2023. Association for Computing Machinery.

[32] Mohamed Elgendi. On the analysis of fingertip photoplethysmogram signals. *Curr. Cardiol. Rev.*, 8(1):14–25, February 2012.

[33] Yeongyeon Na, Minje Park, Yunwon Tae, and Sunghoon Joo. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram, 2024.

[34] Donglai Wei, Joseph J. Lim, Andrew Zisserman, and William T. Freeman. Learning and using the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[35] Bryan Gopal, Ryan Han, Gautham Raghupathi, Andrew Ng, Geoff Tison, and Pranav Rajpurkar. 3kg: Contrastive learning of 12-lead electrocardiograms using physiologically-inspired augmentations. In *Machine Learning for Health*, pages 156–167. PMLR, 2021.

[36] Mostafa Neo Mohsenvand, Mohammad Rasool Izadi, and Pattie Maes. Contrastive representation learning for electroencephalogram classification. In Emily Alsentzer, Matthew B. A. McDermott, Fabian Falck, Suproteem K. Sarkar, Subhrajit Roy, and Stephanie L. Hyland, editors, *Proceedings of the Machine Learning for Health NeurIPS Workshop*, volume 136 of *Proceedings of Machine Learning Research*, pages 238–253. PMLR, 11 Dec 2020.

[37] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

[38] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15745–15753, 2020.

[39] Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

[40] Joseph Y Cheng, Erdrin Azemi, Hanlin Goh, Kaan E Dogrusoz, and Cuneyt O Tuzel. Subject-aware contrastive learning for biosignals, December 2 2021. US Patent App. 17/326,098.

[41] Nathaniel Diamant, Erik Reinertsen, Steven Song, Aaron D. Aguirre, Collin M. Stultz, and Puneet Batra. Patient contrastive learning: A performant, expressive, and practical approach to electrocardiogram modeling. *PLOS Computational Biology*, 18(2):e1009862, February 2022.

[42] Dani Kiyasseh, Tingting Zhu, and David A Clifton. Clocs: Contrastive learning of cardiac signals across space, time, and patients. In *International Conference on Machine Learning*, pages 5606–5615. PMLR, 2021.

[43] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. *CoRR*, abs/2104.14548, 2021.

[44] Tian Bai, Shanshan Zhang, Brian L. Egleston, and Slobodan Vucetic. Interpretable representation learning for healthcare via capturing disease progression through time. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 43–51, New York, NY, USA, 2018. Association for Computing Machinery.

[45] Ying An, Kun Tang, and Jianxin Wang. Time-aware multi-type data fusion representation learning framework for risk prediction of cardiovascular diseases. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(6):3725–3734, 2021.

[46] Monica N Agrawal, Hunter Lang, Michael Offin, Lior Gazit, and David Sontag. Leveraging time irreversibility with order-contrastive pre-training. In *International Conference on Artificial Intelligence and Statistics*, pages 2330–2353. PMLR, 2022.

[47] Qingyu Zhao, Zixuan Liu, Ehsan Adeli, and Kilian M. Pohl. Longitudinal self-supervised learning. *Medical Image Analysis*, 71:102051, 2021.

[48] Sana Tonekaboni, Danny Eytan, and Anna Goldenberg. Unsupervised representation learning for time series with temporal neighborhood coding. In *International Conference on Learning Representations*, 2021.

[49] Hugo Yèche, Gideon Dresdner, Francesco Locatello, Matthias Hüser, and Gunnar Rätsch. Neighborhood contrastive learning applied to online patient monitoring. In *International Conference on Machine Learning*, pages 11964–11974. PMLR, 2021.

[50] Siyi Tang, Jared A Dunnmon, Qu Liangqiong, Khaled K Saab, Tina Baykaner, Christopher Lee-Messer, and Daniel L Rubin. Modeling multivariate biosignals with graph neural networks and structured state space models. In *Conference on Health, Inference, and Learning*, pages 50–71. PMLR, 2023.

[51] Tania Pereira, Nate Tran, Kais Gadhoumi, Michele M Pelter, Duc H Do, Randall J Lee, Rene Colorado, Karl Meisel, and Xiao Hu. Photoplethysmography based atrial fibrillation detection: a review. *npj Digital Medicine*, 3(1):3, January 2020.

[52] Kevin G Montero Quispe, Daniel MS Utyiama, Eulanda M Dos Santos, Horácio ABF Oliveira, and Eduardo JP Souto. Applying self-supervised representation learning for emotion recognition using physiological signals. *Sensors*, 22(23):9102, 2022.

[53] Federico Del Pup and Manfredo Atzori. Applications of self-supervised learning to biomedical signals: where are we now. *TechRxiv*, 2023.

[54] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *Advances in Neural Information Processing Systems*, 32, 2019.

[55] Russell Katz. Biomarkers and surrogate markers: an fda perspective. *NeuroRx*, 1(2):189–195, April 2004.

[56] Natallia Gray, Gabriel Picone, Frank Sloan, and Arseniy Yashkin. Relation between bmi and diabetes mellitus and its complications among us older adults. *Southern Medical Journal*, 108(1):29–36, 2015.

[57] World Health Organization et al. Use of glycated haemoglobin (hba1c) in diagnosis of diabetes mellitus: abbreviated report of a who consultation. Technical report, World Health Organization, 2011.

[58] R J Shephard, C Allen, A J Benade, C T Davies, P E Di Prampero, R Hedman, J E Merriman, K Myhre, and R Simmons. The maximum oxygen intake. an international reference standard of cardiorespiratory fitness. *Bull. World Health Organ.*, 38(5):757–764, 1968.

[59] Kyle Mandsager, Serge Harb, Paul Cremer, Dermot Phelan, Steven E Nissen, and Wael Jaber. Association of cardiorespiratory fitness with long-term mortality among adults undergoing exercise treadmill testing. *JAMA Netw. Open*, 1(6):e183605, October 2018.

[60] Ralph B D'Agostino, Sr, Ramachandran S Vasan, Michael J Pencina, Philip A Wolf, Mark Cobain, Joseph M Massaro, and William B Kannel. General cardiovascular risk profile for use in primary care: the framingham heart study. *Circulation*, 117(6):743–753, February 2008.

[61] Apple. Estimating sleep stages from apple watch. Technical report, Apple, September 2023.

[62] American Diabetes Association Professional Practice Committee. 2. classification and diagnosis of diabetes: Standards of medical care in diabetes–2022. *Diabetes Care*, 45(Supplement_1):S17–s38, 12 2021.

[63] Leonard A Kaminsky, Ross Arena, and Jonathan Myers. Reference standards for cardiorespiratory fitness measured with cardiopulmonary exercise testing: data from the fitness registry and the importance of exercise national database. In *Mayo Clinic Proceedings*, volume 90, pages 1515–1523. Elsevier, 2015.

[64] Connor B. Weir and Arif Jan. *BMI Classification Percentile And Cut Off Points*. StatPearls Publishing, Treasure Island (FL), 2022.

[65] Donna K Arnett, Roger S Blumenthal, Michelle A Albert, Andrew B Buroker, Zachary D Goldberger, Ellen J Hahn, Cheryl Dennison Himmelfarb, Amit Khera, Donald Lloyd-Jones, J William McEvoy, et al. 2019 acc/aha guideline on the primary prevention of cardiovascular disease: a report of the american college of cardiology/american heart association task force on clinical practice guidelines. *Circulation*, 140(11):e596–e646, 2019.

[66] Carlo Alberto Barbano, Benoit Dufumier, Edouard Duchesnay, Marco Grangetto, and Pietro Gori. Contrastive learning for regression in multi-site brain age prediction. In *International Symposium on Biomedical Imaging (ISBI)*, 2023.

[67] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, 2019.

[68] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8765–8775. Curran Associates, Inc., 2020.

Appendices

A Motivations, goals, and proposed advancements.

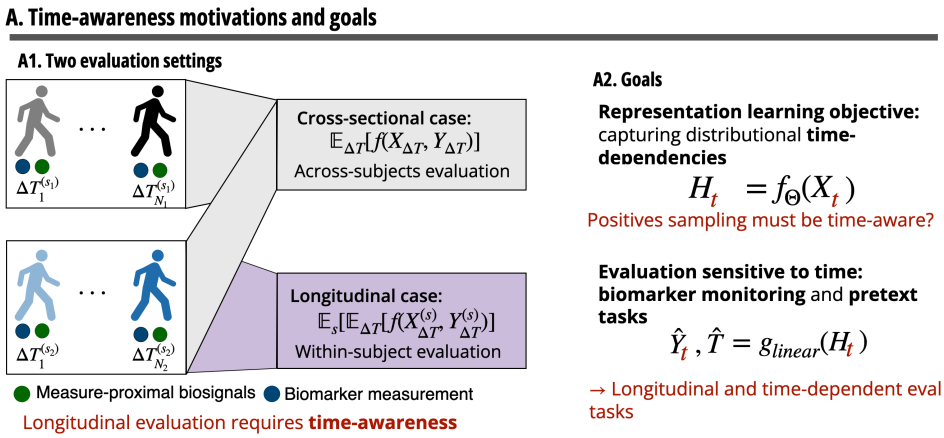

Figure 2: **Motivations for the time-awareness desiderata and proposed advancements.** A1) Illustration of the within-subject versus across-subjects evaluation. A2) Objectives and proposed solutions presented in this work.

B Related Work

B.1 Self-supervised representation learning of biosignals

Large-scale datasets, together with the recent advances of SSL algorithms, have shown significant advances in domains such as natural language processing [24, 25], computer vision [18, 26], and speech [27], and have led to promising foundation models for physiological time series trained without labels [12, 4, 28, 29]. SSL has thus emerged as the dominant paradigm for learning physiological time series representations, outperforming hand-crafted morphological features or supervised approaches on numerous downstream medical tasks [30, 31, 32, 12, 10]. The taxonomy of SSL algorithms often presents joint-embedding architectures and pretext task learning as prominent paradigms, and both have been used for biosignal representation learning. Pretext tasks include signal reconstruction [33, 31], temporal context prediction such as relative positioning and temporal un-shuffling [30], or prediction of the direction of time directly [30, 34].

Joint-embedding architectures proved efficient among self-supervised approaches, with different strategies developed to prevent representation collapse, including the use of negative in-batch samples or within memory banks in contrastive settings [18, 35, 36], momentum training or batch normalization in non-contrastive settings [37], gradient-blocking [38], or other forms of regularization [39], including for capturing the inter-subject variability of biosignals [7, 40, 41]. Following its recent success, this work focus on encouraging time-sensitivity for joint embedding architectures.

This type of architecture often relies on the definition of positive pairs of samples, which specify the type of invariances that the model is enforced to learn [42]. In the visual domain, using two augmented views of data as positive pairs has proven effective for learning representations invariant to small deformations [18]. This approach was later extended to mine positive pairs from neareast-neighbors examples within the training data [43]. For time series data, recent studies have shown the success of instead sampling two segments of data from the same participant to create a positive pair, combined with a well-tailored augmentation module [12] that further improves the ability to capture pertinent inter-subject variability [7, 12, 40, 41, 31, 42]. If the goal is to make cross-sectional predictions, e.g., to differentiate between diseased and healthy subjects, then such an approach may be sufficient. While well-adapted to capture *across*-subject differences in physiology, enforcing such strong subject-invariances contradicts the leitmotif of this work to also account for temporal dependencies *within*-subjects. We hypothesize that capturing such intra-subject variability is crucial to

making good predictions in health status over time. For instance, biosignals from the same participant might look quite different if sampled years apart and a major medical event, such as a heart attack, occurred between them.

B.2 Representation learning for longitudinal biomedical data

Recent work used electronic health record data to capture disease progression patterns from sequential visit data, with labels to learn time-aware decay functions and attention mechanisms to learn representations of each visit [44, 45]. Alternatively, the temporal irreversibility of disease progression has been explored to design an order-contrastive self-supervised task, where the objective is to predict whether pairs of longitudinal time segments are in the correct order [46]. Longitudinal self-supervised learning has also been explored for MRI brain neuroimaging via factor disentanglement to align the image representations over time with the *brain age* [47].

As for biosignals, temporal dynamics have mostly been exploited *within* the time series, where it is generally assumed that temporally close segments must share similarities and context information. This has been used to design pretext tasks [30], or contrastive tasks that define positives views as temporally close segments [42]. The approach in [48] defines positive pairs using an anchor-dependent temporal neighborhood, parametrized by a Gaussian kernel centered on the anchor timestamps, and a range estimated using the Augmented Dickey-Fuller (ADF) statistical test that enforces the signal stationarity within the neighborhood. Running the statistical test on every anchor window makes it prohibitive for datasets substantially larger than those used in their work, which were limited to fewer than 30 participants and lacked longitudinal data. Most similar to this work in spirit, recent research proposed a neighborhood-aware loss composed of two terms: a first term encourages representations from the same neighborhood (e.g., subject) to be close and from different neighborhoods to be far apart, while a second term encourages representation diversity for all data within a neighborhood [49]. While the authors consider hard neighbor assignments, our work uses a soft definition of temporal neighborhood sampled using Gaussian kernels.

B.3 Health monitoring from biosignals

SSL is particularly suited for label-scarce medical settings, and has has been exploited to create representations of physiological signals like PPG predictive of seizure detection [50], left ventricular hypertrophy and atrial fibrillation [51], and for the recognition of activity and emotion [10, 52]. See [53] for a recent review on the use of SSL for biosignals. This work extends the evaluation of biosignal representations to time-varying medical targets in longitudinal settings, and aims at computing versatile representations predictive across biomarkers and over time.

C Implementation details

C.1 Model architecture and training configuration

The backbone encoder was an EfficientNet-style 1D convolutional neural network with 16 mobile-inverted bottleneck blocks with squeeze-and-excitation [16]. Previous work not published here showed that EfficientNets provide the best balance between performance and parameter size in comparison to ResNets and Transformers. The architecture of the EfficientNet is depicted in Figure 3.

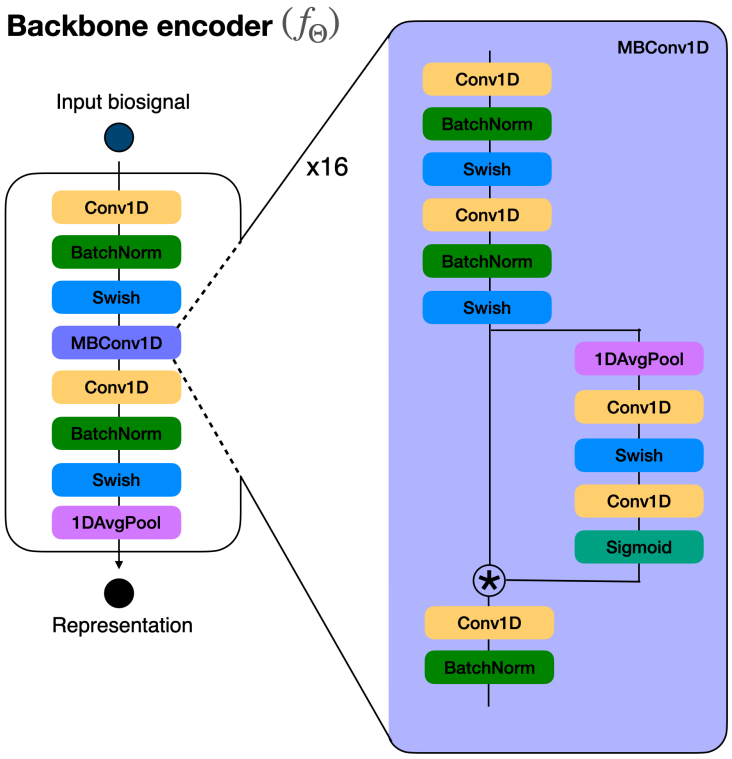

Figure 3: **Architecture of our EfficientNet-style encoder, adapted for multivariate time series inputs from [16].** Conv1D: Convolutional block; BatchNorm: batch normalization block; Swish: Swish activation block; MBConv1D: Mobile inverted bottleneck block; 1DAvgPool: Average pooling; Sigmoid: sigmoid activation block; asterix (*): element-wise multiplication.

We used representations with 256 dimensions, computed after the backbone (the projection head is used for training only). The encoder had 3.3M parameters. The projection head was a multi-layer perceptron with one hidden layer of 1024 units, and a dropout rate of $p = 0.1$, mapping the 256-dimensional embeddings to a 128-dimensional representation subspace where the loss is calculated. In practice, we added a small offset $\epsilon = 10^{-6}$ to the densities $s_T(d_j)$ and $s_H(\theta_j)$ before normalizing, to avoid collapse in the number of potential positive pairs for cases where an anchor might have very few candidates close in time and/or time of day.

For a batch of B pairs of positive pairs $(h_1^1, h_2^1), ..., (h_1^B, h_2^B)$, the loss is computed as $L_{InfoNCE}^{contrastive} = \frac{1}{2}(L_{1,2} + L_{2,1})$, with

$$L_{u,v} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(<h_u^i|h_v^i>_\tau)}{\sum_{j=1}^{B}\exp(<h_u^i|h_v^j>_\tau)}, \tag{1}$$

using the $\mathcal{L}_2$-normalized temperature-scaled cosine similarity $< a|b >_\tau = \frac{a^T b}{\tau\|a\|\|b\|}$. Intuitively, this enforces representations of positive pairs to be attracted in the latent space by maximizing their mutual information, while repulsing the representations of the negative contrastive samples within

the batch [54]. For a given anchor, negative samples consist of all other samples in the batch from different subjects, and no hard negative mining was used. The temperature value for the InfoNCE loss was set to $\tau = 0.04$, the batch size was set to $B = 256$, the initial learning rate was 0.001, and we used step learning rate schedule for faster convergence, with a step size of 200 and $\gamma = 0.5$. Positive segments were redrawn during each batch of training to create the positive pairs. We trained the joint embedding architecture in a student-teacher training scheme, where one side is updated with back-propagation (student), and the other side (teacher) is an exponential moving average of the student side [12]; the student side was updated using Adam optimizer with default PyTorch parameters, distributed across 32 A100 GPUs (4 nodes, each with 8 A100 GPUs, 1TB of RAM, and 96 CPUs), and the momentum update for the teacher side was set to 0.99. Models generally finished training (i.e. validation loss converged) after about 6 days of training.

C.2 Time-dependent sampling of the positive pairs

In our experiments, we compare the standard time-independent uniform sampling, that is, $s = s_{uniform} \hookrightarrow \mathcal{U}$, with a sampling dependent on (i) time (days), (ii) hour of day, or (iii) both of these. For (i), we define $s_T \hookrightarrow \mathcal{N}(d_i, \sigma_T)$, and test $\sigma_T \in [14, 42, 126]$ days in our experiments. To sample a positive pair, we compute a vector of weights p_T^i by evaluating the density of each $s_T(d_j)$ for $j \neq i$, then sample from the resulting multinomial after normalizing. For (ii), we define $s_H \hookrightarrow \mathcal{N}(\theta_i, \sigma_H)$, construct a similar weights vector $p_H(\theta_j)$ and test $\sigma_H \in [1.5, 3, 6]$ hours. To appropriately handle the constraint $\theta_i \in [0, 24)$, in practice we sample from a circular normal (von Mises) distribution rather than a standard Gaussian, with $\kappa = 1/\sigma_H^2$. For (iii), we use a mixture of both components $p_{TH} = p_T * p_H$, ensuring the sampled pair is close both in time and hour of day. For clarity, we use time versus hour-based to name both types of sampling. Figure 4 illustrates the uniform time-independent and the time-dependent sampling procedures used to create the positive pairs. Algorithm 1 below provides code for this, and see Algorithm 1 in the appendix of [12] for pseudocode for the overall pre-training framework.

Algorithm 1: Code for our proposed time-dependent positive pair sampling

```
# inputs
#  a_date:  anchor segment date, in [0,T] (days since first observation at time 0)
#  a_hour:  anchor segment hour of day, in [0,24)
#  c_dates:  np array of dates for candidate segments, in [0,T]
#  c_hours:  np array of hours of day for candidate segments, in [0,24)
#  sigma_T: kernel for day sampling.  if 0, do not use
#  sigma_H: kernel for hour sampling.  if 0, do not use
#  eps:  small offset to avoid collapse of weight vectors, default 1E-6
# if both sigma_T and sigma_H are 0, this turns into uniform sampling baseline with equal weights
n_candidates = len(c_dates)
# convert hours to angles
a_angle = 2 * np.pi * a_hour / 24
c_angles = 2 * np.pi * c_hours / 24
# get probabilities for time/day sampling
if sigma_T > 0:
    pos_seg_prob_t = scipy.stats.norm.pdf(c_dates - a_date, loc=0, scale=sigma_T)
else:
    # uniform
    pos_seg_prob_t = np.ones(n_candidates) / n_candidates
# get probabilities for hour sampling
if sigma_H > 0:
    # convert units of sigma_H from hour to angle
    std_dev_radians = (sigma_H / 24) * 2 * np.pi
    # scale param for circular gaussian pdf
    kappa_radians = 1 / (std_dev_radians) ** 2
    pos_seg_prob_h = scipy.stats.vonmises.pdf(0, kappa=kappa_radians, loc=c_angles - a_angle)
else:
    # uniform
    pos_seg_prob_h = np.ones(n_candidates) / n_candidates
# add small offset and renormalize to smooth probabilities
pos_seg_prob_t += eps
pos_seg_prob_t /= pos_seg_prob_t.sum()
pos_seg_prob_h += eps
pos_seg_prob_h /= pos_seg_prob_h.sum()
# compute mixture of sampling probabilities
pos_seg_prob = pos_seg_prob_t * pos_seg_prob_h
pos_seg_prob /= pos_seg_prob.sum()
# draw positive pair from sampling probabilities
pos_seg_ind = np.random.choice(np.arange(n_candidates), p=pos_seg_prob)
return pos_seg_ind
```

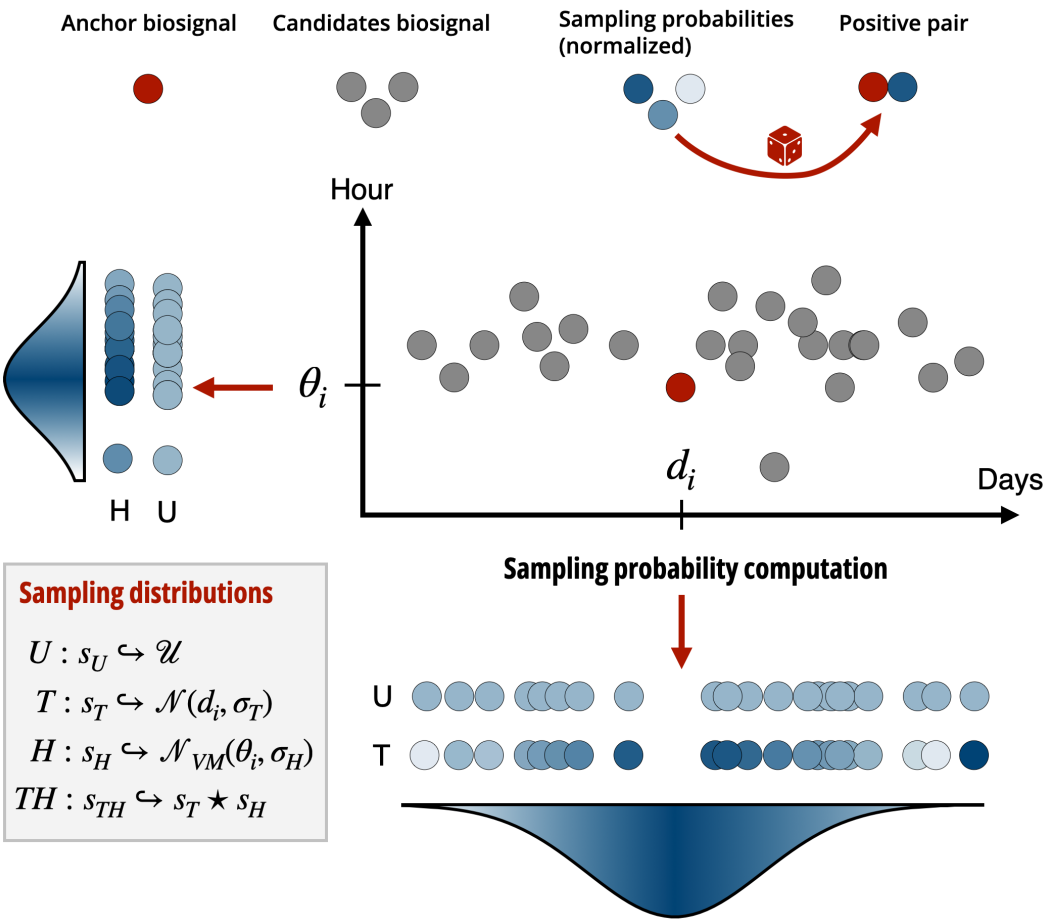

Figure 4: **Illustration of the time-dependent and time-independent sampling strategies used to form a pair of biosignals *positively* associated.** At runtime when building the batch of training data, we compute for each anchor biosignal a weight vector (of probabilities) used to draw a positive example from the set of candidates. The set of candidates consists of the same participant's biosignals except the anchor segment. During training, the InfoNCE contrastive loss optimization enforces the model to maximize the mutual information between the positive pair representations. The distributions $\mathcal{U}, \mathcal{N}$ and $\mathcal{N}_{VM}$ denote the uniform, Gaussian, and von Misses distributions with $\kappa = 1/\sigma_H^2$.

D Dataset presentation

D.1 Photoplethysmogram (PPG) collection and pre-processing.

AHMS is an ongoing digital research study exploring the links between physical activity and cardiovascular health [15]. It is sponsored by Apple and conducted in partnership with the American Heart Association and Brigham and Women's Hospital. To be eligible for the study, participants must be at least 18 years of age (21 in some locations), reside in the United States, have access to an Apple Watch, and provide informed consent electronically in the Apple Research app. Apple Watch passively records green-light PPG signals during low-motion periods throughout the day using light-emitting and light-sensitive diodes. Recorded PPG signals are sampled at 64Hz or 256Hz for 60 seconds, and consist of four separate optical channels, each associated with different spatial combinations of transmitting and receiving diodes. PPG preprocessing included dark subtraction (to remove ambient light), followed by band-pass filtering, resampling to 64Hz if needed, and temporal channel-wise z-scoring for each segment.

D.2 Outcomes of interest

Often, cardiovascular and metabolic health is quantified using risk factors or biomarkers that trend with disease severity [55]. In this work, we use BMI and hemoglobin A1C as measures of metabolic health [56, 57], estimated VO2Max for assessing cardiopulmonary fitness [58, 59], and the Framingham risk score for cardiovascular health [60]. The Framingham risk score is a sex-specific multi-variable estimate of the risk of developing cardiovascular disease events (coronary, cerebrovascular, peripheral arterial disease or heart failure) within 10 years. It is computed based on sex, age, LDL cholesterol, HDL cholesterol, blood pressure, and diabetes, smoking, and hypertension treatment status. To the best of our knowledge, this is the first study to use prediction of the hour of day from biosignal representations as a proxy to quantify sensitivity to circadian rhythms, and to predict cardiovascular risk directly from PPG.

Dataset	**# Subjs**	**# PPGs**	**# Subj-90-day bins**	**% Class**
Time-of-day	222,493	3.55M	-	40.2% day, 59.8% night
Sleep-wake	49,377	866K	-	3.6% awake, 96.4% sleep
Framingham	3,667	2.42M	23,449	39.5% low, 16.3% borderline, 33.0% intermediate, 11.2% high
VO2Max	176,575	8.40M	783,513	32.5% low, 28.8% low-med, 22.3% med-high, 15.8% high
A1C	12,045	1.77M	24,612	52.3% normal, 27.2% prediabetes, 14.5% diabetes, 5.3% severe diabetes
BMI	77,617	5.68M	293,112	0.7% underweight, 27.8% normal, 36.7% overweight, 34.7% obese

Table 2: **Number of participants and PPG segments used in each task.** We use the data from 80% of the participants for training downstream models, and 20% to perform the final test evaluation, which is what we report. We also report outcome class percentages per dataset. The time-of-day and sleep-wake datasets are per-segment (each PPG has its own label, and models were fit and evaluated per PPG segment, so class percentages are reported per segment). The Framingham, VO2max, A1C, and BMI datasets are per subject-90-day-periods, so class percentages are reported per subject-90-day-period.

E Evaluation procedure

Datasets creation The binary sleep-wake dataset was created for users with sleep staging enabled on their Apple Watch [61]. We randomly subsampled 20 PPG segments recorded during overnight sleep stages. Apple Watch's sleep stages are categorized as Awake, Deep, Core and REM, are available to the users on the Health App and are derived from an accelerometer-based algorithm from Apple Watch [61]. For our wake-sleep dataset, we assigned the Awake stage to the Wake class, and any of Deep, Core and REM stages to the Sleep class.

To assemble the longitudinal datasets, we started by taking all PPGs that were recorded within 30 days of an A1C or Framingham score measurement, or within 10 days of a BMI or VO2max measurement. We then subsampled PPGs depending on the outcome: for Framingham risk, we subsampled to a max of 100 segments per non-overlapping 30 day period; for A1C, we subsampled to a max of 3 segments per day; for BMI, we subsampled to a max of 10 segments per non-overlapping 30 day period; and for VO2max, we subsampled to a max of 5 segments per non-overlapping 30 day period. Our base units of observation for evaluation are non-overlapping 90-day periods of time per subject. For instance, a subject with 360 days between their first and last PPGs (that are also proximal to the outcome variable) would contribute 4 such 90-day periods. Each 90-day period is labeled according to the most common class label per outcome.

Task labels definition per clinical outcome A1C was binned into the usual categories of normal ($< 5.7\%$), prediabetes ($\geq 5.7\%$ & $< 6.5\%$), diabetes ($\geq 6.5\%$ & $< 8.0\%$), and severe diabetes ($\geq 8.0\%$) [62]. We used the age and sex of the participants to map individual VO2max values into population quartiles per age-sex bin (i.e. 0-25% is "low", 25-50% is "low-medium", 50-75% is "medium-high", 75-100% is "high") using reported values from [63]; see Appendix H for exact values. Subjects without a reported age or sex were excluded from the VO2max dataset. For BMI, we used standard categories of underweight (< 18.5), normal (≥ 18.5 & < 25), overweight (≥ 25 & < 30), and obese (≥ 30) [64]. Standard risk thresholds were used for the Framingham score: low ($< 5\%$), borderline ($\geq 5\%$ & $< 7.5\%$), intermediate ($\geq 7.5\%$ & $< 20\%$), and high ($> 20\%$) [65].

For BMI, different validated cutoffs were used in Asian and South-Asian populations compared to White, Hispanic, and Black populations to define normal, overweight, obsese, etc bins [64]. We excluded participants who did not self-report an ethnicity or biological sex.

Classes definitions for both low-and-high-contrast classification tasks are reported in the explanatory Figure 5.

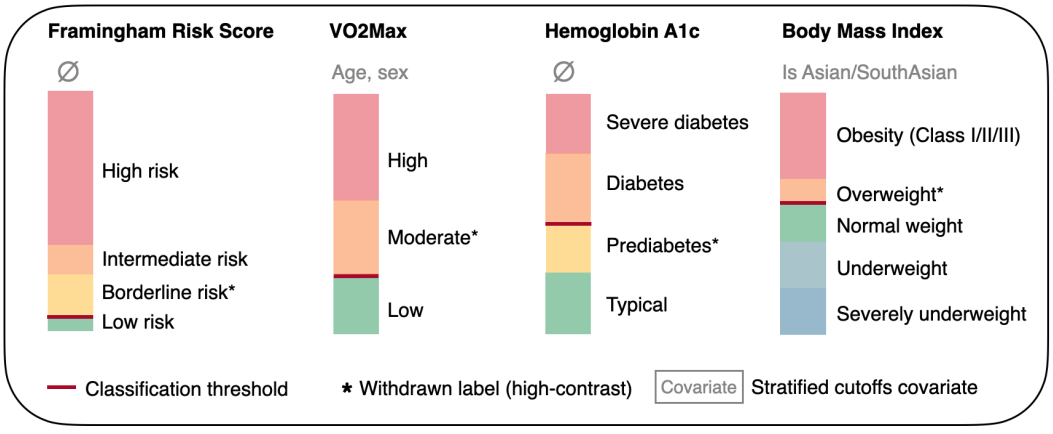

Figure 5: **Biomarkers diagnosis labels and classification cutoffs.** For all outcome, diagnosis groups above the red bar constitute the positive class of the classification task. In high-contrast scenario, the diagnosis group with a star is withdrawn.

Tasks performances computation. For both direct and indirect evaluations, we used linear probing with ridge-penalized linear models on the representations produced by the trained biosignal encoders (without the projection head) [20]. The models for the time-of-day dataset and sleep-wake datasets

were trained and evaluated per PPG segment. For the other 4 datasets, we first mean aggregated the embeddings within each 90-day period for each subject, before training and evaluating models on a per subject-90-day period basis. For each task, we choose the best kernel combinations for the three proposed positive pair sampling schemes, so different kernels (and hence models) are reported for each task. We quantify performance with the area under the ROC curve (AUC - which is also equivalent to the probability that a model ranks a random positive example higher than a random negative). On all tasks, we evaluate in both a *within-subject* as well as *across-subject* manner.

We evaluate *within-subject* by only making comparisons between segments from the same subject for the time-of-day and sleep-wake tasks, or between 90-day periods from the same subject for the biomarker monitioring tasks. To compute an AUC like this, we count the proportion of times that the model correctly ranked the positive class as higher than the negative class *from the same subject* – that is, we never include in this calculation a positive segment or period from subject i compared with a negative segment or period from subject $j \neq i$. This within-subject AUC is also equivalent to computing an AUC separately within each subject, and taking a weighted average across subjects (weighted by the number of comparisons made per subject). See Algorithm 2 for code that implements this metric.

We also evaluate in an *across-subject* manner, to demonstrate how this more common cross-sectional strategy can lead to very different conclusions than the previous *within-subject* evaluation. To do this, we stack up all labels and model scores for all segments or 90-day periods across all subjects, ignoring subject identity. Unsurprisingly, this form of evaluation always yields better results, in part because our representations were trained to explicitly separate subjects. It is thus much easier for a model to correctly rank different people with very different physiological states than it is to correctly identify changes within a single individual.

Algorithm 2: Code for computing the within-subject AUROC binary classification metric, so that rankings are only made between time periods from the same subject.

```
# inputs
# y_true:np array of binary labels (1's and 0's) for each time period
# y_score:np array of model scores (assumed to be in [0,1]) for each time period
# y_id:np array of subject identifiers, denoting which subject each time period came from
df = pd.DataFrame({"y_id":y_id, "y_true":y_true, "y_score":y_score})
# compute auc separately for each subject, and track number of positive and negative periods
metricdf = df.groupby("y_id").apply(
    lambda x:  pd.Series({
        "n_pos":  x["y_true"].sum(),
        "n_neg":  x.shape[0] - x["y_true"].sum(),
        # only compute auc for this subject if at least 1 positive and negative period
        "auc":  (skmetrics.roc_auc_score(x["y_true"], x["y_score"]) if x["y_true"].sum() > 0 and
         x["y_true"].sum() < x.shape[0] else 0)
    })
)
# total number of correct within-subject comparisons made
within_subj_correct = (metricdf["auc"] * metricdf["n_pos"] * metricdf["n_neg"]).sum()
# total number of within-subject comparisons made
within_subj_total = (metricdf["n_pos"] * metricdf["n_neg"]).sum()
# within-subject auc is not defined if no within-subject comparisons were made!
if within_subj_total == 0:
    within_subj_roc_auc_score = np.nan
else:
    within_subj_roc_auc_score = within_subj_correct / within_subj_total
return within_subj_roc_auc_score
```

F Across-subjects linear classification performances.

Dataset	Task	Positive pairs sampling			
		U	T	H	TH
Time-of-day		0.876 (.000)	0.876 (.000)	0.889 (.000)	**0.889** (.000)
Sleep-Wake		0.843 (.000)	0.860 (.000)	**0.870** (.000)	0.867 (.000)
A1C	LC	0.843 (.001)	0.843 (.001)	**0.845** (.001)	0.842 (.001)
	HC	0.895 (.001)	0.895 (.001)	**0.899** (.001)	0.892 (.001)
BMI	LC	**0.912** (.000)	0.911 (.000)	0.911 (.000)	0.910 (.000)
	HC	0.971 (.000)	0.971 (.000)	0.971 (.000)	**0.971** (.000)
VO2Max	LC	0.884 (.000)	**0.884** (.000)	0.884 (.000)	0.883 (.000)
	HC	0.961 (.000)	**0.961** (.000)	0.961 (.000)	0.960 (.000)
Framingham	LC	0.922 (.001)	0.921 (.001)	0.914 (.001)	**0.922** (.001)
	HC	0 .940 (.001)	**0.940** (.001)	0.936 (.001)	0.938 (.001)

Table 3: **Across-subject performance of PPG representations using proposed time-aware sampling of positive pairs.** LC and HC refer to low and high-contrast task versions per outcome, evaluating ability to correctly rank 90-day periods across all subjects. “U" denotes baseline time-invariant uniform sampling of positive pairs, and “T", “H", and “TH" correspond to our proposed positive-pair sampling that is sensitive to number of days, hour of day, or both. Standard error of the means over 200 bootstrap-resampled test sets are shown in parenthesis below each AUC. Model with highest AUC point estimate in bold, models non-inferior to the best model based on a paired one-sided Wilcoxon signed-rank test using bootstrapped AUCs are underlined ($p > .05$).

G Discussion, limitations, and future work

We view this work as only beginning to scratch the surface in developing more time-aware methods for biosignal representation learning, and there are many potential avenues to further improve this work. In some cases, we suspect that our proposed sampling strategies may increase positive pair redundancy in low-density regions of the sampling space, which may hurt the learned representations. This could be improved by more careful sampling strategies that attempt to reduce this redundancy when an anchor segment has few potential positive pair candidates. A minor avenue for improvement would be to better tune the training hyperparameters for time-aware models; we used settings for the baseline uniform sampling of positive pairs from [12], and these were fixed across the proposed more time-aware models. An additional limitation is our reliance on self-reported or clinical health record-based biomarkers as time-varying labels to predict given that inaccuracies in these labels—e.g., errant survey responses or incorrectly coded health records—could frustrate the development and validation of such models. Further investigation is also warranted to try to better understand the physiological underpinnings of our results, perhaps by visualizing the learned PPG representations from individual subjects over time.

The extent to which subtle biosignals variations, reflecting a changing health status over time, can be captured using time-dependent sampling might be limited in practice, and outcome-dependent. This suggests the need for practitioners to test different options for making a baseline model architecture more time-aware, until potentially one is found with better performance for the task at hand. We believe that even better time-sensitivity can be achieved than what we propose in this work by combining multiple approaches. For instance, one could make use of even larger longitudinal datasets with enhanced temporal granularity, and combine a time-aware sampling strategy as proposed in this work with additional fine-tuning or personalization of the encoder for a given application. Furthermore, the inclusion of other types of metadata in positive pair selection, such as is [66], might also be beneficial – e.g. going beyond just hour of day and time between biosignals, and using other contexts that may be available such as day-of-week effects (e.g. weekday vs weekend), and whether someone recently exercised, was sick, or had poor sleep, all of which could manifest in a biosignal like PPG. There are undoubtedly other approaches that can also better handle the *sampling* irregularity and *bias* [67, 68], missing data, and label noise that are common in large longitudinal wearables datasets, as well as techniques to improve the overall data efficiency, interpretability, generalizability, and robustness of learned representations from biosignals.

H VO2max demographic percentiles

TABLE 3. Sex-Specific Percentiles for CRF From Treadmill Exercise Tests With Measured $\dot{V}O_{2max}$ Obtained From FRIEND and Predicted $\dot{V}O_{2max}$ ($mLO_2 \cdot kg^{-1} \cdot min^{-1}$) Reported by the Cooper Clinic[13,a,b]

	Percentile						
Age group (y)	5th	10th	25th	50th	75th	90th	95th
Men from FRIEND[c]							
20-29	29.0	32.1	40.1	48.0	55.2	61.8	66.3
30-39	27.2	30.2	35.9	42.4	49.2	56.5	59.8
40-49	24.2	26.8	31.9	37.8	45.0	52.1	55.6
50-59	20.9	22.8	27.1	32.6	39.7	45.6	50.7
60-69	17.4	19.8	23.7	28.2	34.5	40.3	43.0
70-79	16.3	17.1	20.4	24.4	30.4	36.6	39.7
Women from FRIEND[c]							
20-29	21.7	23.9	30.5	37.6	44.7	51.3	56.0
30-39	19.0	20.9	25.3	30.2	36.1	41.4	45.8
40-49	17.0	18.8	22.1	26.7	32.4	38.4	41.7
50-59	16.0	17.3	19.9	23.4	27.6	32.0	35.9
60-69	13.4	14.6	17.2	20.0	23.8	27.0	29.4
70-79	13.1	13.6	15.6	18.3	20.8	23.1	24.1

[a]CRF = cardiorespiratory fitness; CPX = cardiopulmonary exercise testing; FRIEND = Fitness Registry and the Importance of Exercise National Database; $\dot{V}O_{2max}$ = maximal oxygen uptake.
[b]All patients are considered free of known cardiovascular disease.
[c]The FRIEND CRF data were measured with CPX.

Figure 6: **VO2max percentiles per demographic subgroup, from the FRIEND study [63].**

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Claims are made explicit at the end of the abstract and of the introduction, with limitations highlighted in the discussion section. Justifications of the claims and contributions are presented throughout the method and result sections.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations of our work (increase in positive pairs redundancy in low-density regions, label noise, potential limited gain of time-awareness in practice) are identified and accompanied with potential solutions in the discussion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All relevant details for reproducibility (foundation model training, time-based sampling of the positive pairs at runtime, the evaluation protocol with datasets and tasks presentations) should be mentioned in the paper. Interested parties can directly contact the authors for code snippets or additional details.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The dataset is not able to be publicly shared due to the specifics of the informed consent, but it has been extensively used in prior publications. We are unable to release code in order to protect subject confidentiality due to specific language in the study protocol and informed consent. Interested parties can directly contact the authors for code snippets or additional details.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Implementation details regarding the training/testing models and procedures are detailed in the method section 2 and appendix C.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Results tables provide standard error of the AUC means over 200 bootstrap-resampled test sets. Models non-inferior to the best model for each task are identified based on a paired one-sided Wilcoxon signed-rank test using bootstrapped AUCs.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Detailed information on the compute resources for training our models can be found in appendix C.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted as part of this work conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: We do not contain any explicit discussion of this topic. As far as we can tell, no specific potential negative societal impacts have been identified following this research. Any application of machine learning in the health space always poses challenges related to things such as bias, fairness, etc but our work is primarily methodological and not yet close enough to a real deployable health application that we did not emphasize these details. Future work that is more applied that extends ours might consider looking into things such as subgroup fairness metrics to see how well our representations work in different groups.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We are not releasing any models or data, as we are unable to do so due to the specifics of the participant informed consent.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: This paper did not make use of any such existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: This paper does not publicly release any such new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: Our work uses data from an IRB-approved study (Apple Heart and Movement Study), and participants all signed informed consent that allowed for their anonymized data to be shared for research purposes with specific parties.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: The study was approved by the Advarra Central Institutional Review Board, and registered to ClinicalTrials.gov (Identifier: NCT04198194) [15].

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.