DEPTH BEATS BREADTH: SCALING WEARABLE WORLD MODELS OF HUMAN PHYSIOLOGY

Anonymous authors
Paper under double-blind review

ABSTRACT

Most wearable foundation models operate as static encoders, summarizing windows of observed signals into representations for downstream health prediction. We develop an action-conditioned wearable *world model* that forecasts physiological trajectories from personal history and specified future behavior. How to scale such models is unclear: adding users and extending personal histories increase data volume differently, and longer histories also provide more inference context. Using a corpus of 5.4 million users and approximately 2 billion user-days, we study four scaling axes: population size, training context, model capacity, and inference context. Across the configurations studied, longer personal histories improve forecasting more than additional users, with much of the benefit coming from longer inference context. Larger models benefit more from long context, while scaling gains are most pronounced in event-anchored cohorts and composite physiological targets. We evaluate forecasting on 231 target–cohort–horizon combinations spanning 11 primary targets, six cohorts, and horizons of 1–180 days. Frozen model representations also support 15 downstream probes: 10 subject-level health targets, including type 2 diabetes and sleep apnea, and five forecasts of self-reported sickness or menstrual-cycle phase at different horizons. Together, these results provide a recipe for scaling wearable world models and a foundation for anticipating changes in health from longitudinal physiology and behavior.

1 INTRODUCTION

Wearable devices record physiology and behavior passively during everyday life, often over months or years. These measurements complement electronic health records, which provide rich clinical histories but are largely organized around encounters with the healthcare system. Repeated observations between those encounters provide a longitudinal view of an individual's physiology in the context of their daily behavior. This creates an opportunity to build wearable world models. A world model learns a representation of a system's state from observations and predicts future states conditional on actions (Ha & Schmidhuber, 2018; Hafner et al., 2025). For wearable physiology, the observations are a person's physiological and behavioral history, and the action inputs specify future behavior, such as sleep duration, activity level, or alcohol use. Such a model could estimate a person's current physiological state, forecast how it will evolve, and eventually help evaluate how changes in behavior might alter that trajectory.

Wearable foundation models learn representations from biosignals and behavioral summaries for health prediction (Narayanswamy et al., 2025; 2026; Erturk et al., 2025). Wu et al. (2026) characterize the dominant approach as a static-encoder paradigm, in which short observation windows are mapped to representations or predictions, and call for models that use longitudinal context to anticipate health trajectories. Existing generative evaluations include extrapolating minute-resolution sensor signals over minutes to hours (Narayanswamy et al., 2025; 2026). Our focus is instead on using months of personal history to forecast physiological trajectories over days to months, conditional on specified future behaviors. This addresses a central part of the proposed shift toward longitudinal modeling and raises the question of how such models should scale.

To our knowledge, we present the first population-scale wearable world model for forecasting longitudinal physiological trajectories conditional on future behavior (Figure 1). The model combines an encoder that maps daily physiological and behavioral history into latent states with an action-

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.

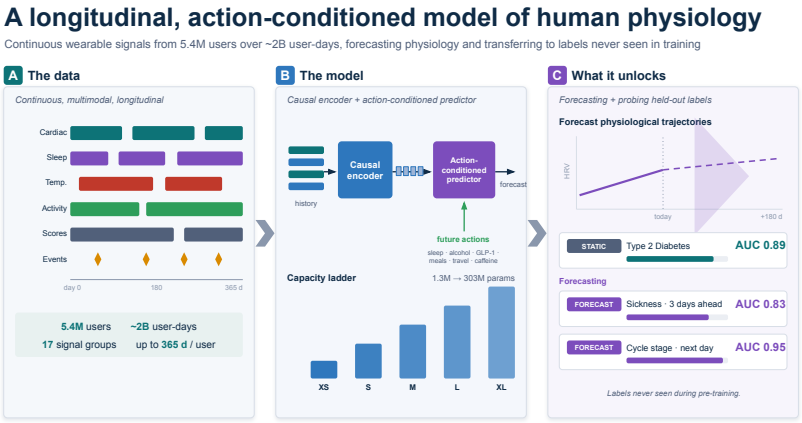

Figure 1: **Overview.** (A) Longitudinal, multimodal wearable records with missing observations and user-reported event tags. (B) A causally masked encoder feeds an action-conditioned predictor of future physiology; we train five model sizes. (C) The model forecasts physiological trajectories, and frozen representations support health-condition detection and event forecasting. Probe labels are not used as pre-training targets.

conditioned predictor of future physiology. We characterize its scaling behavior using a corpus of 5.4 million users and approximately 2 billion user-days—over five times as many user-days as SensorFM's reported maximum of 350 million (Narayanswamy et al., 2026). The model predicts physiological trajectories from daily summaries of past physiology and behavior, conditioned on future action inputs. Our experiments evaluate prediction on observational data; they do not establish that changing an action input recovers the causal effect of changing that behavior. Counterfactual prediction is a longer-term goal that requires additional assumptions and validation.

Scaling studies in language, vision, and time-series forecasting have established relationships among data volume, model capacity, and performance (Kaplan et al., 2020; Hoffmann et al., 2022; Zhai et al., 2022; Shi et al., 2024; Muennighoff et al., 2023). Longitudinal wearable data introduce a further distinction: more data can mean more individuals or more history from each individual. Earlier wearable studies examined subject count and recording duration while holding the model's input window fixed (Narayanswamy et al., 2025). In our setting, additional history can also lengthen the context on which a prediction depends. We therefore distinguish four scaling axes: *breadth*, the number of users; *training depth*, the length of personal history used during training; *capacity*, the number of model parameters; and *inference context*, the history supplied to a trained model at evaluation. This design distinguishes the benefit of training with longer histories from the benefit of providing more history to an already-trained model.

Within the ranges studied, *longer personal history improves forecasting more than additional users* (depth beats breadth), with much of the benefit coming from longer inference context. *Capacity and context interact*: larger models generally gain more from longer histories, but a larger model given seven days of context can underperform a smaller model given a full year. *Scaling gains depend on the cohort and target*: several event-anchored cohorts, particularly postpartum, show larger improvements than the population-wide evaluation, while composite targets benefit more from capacity than individual physiological measurements. Finally, *forecast-trained representations transfer beyond biometric prediction*: frozen model representations support prediction of subject-level health labels, including sleep apnea and type 2 diabetes, and forward-projected representations support forecasts of self-reported sickness and menstrual-cycle phase. These labels are not targets of the biometric forecasting objective. Together, these findings identify which increases in data, capacity, and context improve performance and provide evidence that the learned representations support a broader set of health prediction tasks.

Table 1: **Scale and temporal context of wearable foundation models.** Values are reported in or derived from the cited studies. *Users* counts pre-training participants; *per-user span* describes recording coverage; and *context* is the temporal extent of history supplied to the model at once. *User-days* reports calendar coverage, observed or sampled days, or 24-hour duration equivalents, depending on the study; parenthetical hours report signal or sampled-window duration. Our corpus combines population-scale coverage with year-long context for action-conditioned physiological forecasting. Appendix A.1 details the reporting conventions, conversions, and the distinction between observed and imputed calendar days.

Model	Input representation	Users	User-days	Per-user span	Context	Modeling
Raw-waveform encoders						
Apple PPG/ECG (Abbaspourazad et al., 2024)	PPG / ECG waveform	141K / 107K	16M[†] (0.4M h)	93 d / 23 d	60 s / 30 s	Static, contrastive
Yuan et al. (Yuan et al., 2024)	Tri-axial accel. @ 30 Hz	100K	0.7M (17M h)	∼7 d (single wear)	10 s	Static, multi-task SSL
PaPaGei (Pillai et al., 2025)	PPG waveform (clinical/PSG)	13.5K	2.4K (58K h)	single session	10 s	Static, contrastive
Minute-resolution sensor encoders						
LSM-1 (Narayanswamy et al., 2025)	6 sensors, per-minute	165K	1.7M[‡] (40M h)	≤10 windows / user	5 h	Static, masked recon.
LSM-2 (Xu et al., 2025)	5 sensors, per-minute	60K	3.6M (40M h)	∼59 d	24 h	Static, masked recon.
SensorFM (Narayanswamy et al., 2026)	34 features / 5 sensors, per-min.	5.0M	350M (>2B h)	70 d (10 weeks)	24 h	Static, masked recon.
Aggregated behavioral encoders						
WBM (Erturk et al., 2025)	27 behavioral metrics, hourly	162K	106M (2.5B h)	∼94 weeks	1 week	Static, contrastive
Ours	74 biometrics + behavior, daily	**5.44M**	**1.8B observed** **2.2B calendar**	**1.7M users ≥365 d**	**up to 365 d**	**Dynamic, action-conditioned**

[†]Apple's calendar coverage spans sparse waveform recordings. [‡]LSM-1 reports sampled window-hours converted to 24-hour equivalents; PaPaGei likewise uses a duration equivalent rather than distinct calendar days. For our corpus, 2.2B calendar user-days includes imputed days between observed days; 1.8B user-days have observations.

2 Related Work

Time-series and clinical foundation models. General-purpose models learn forecasting and representations from heterogeneous time-series corpora (Garza et al., 2023; Rasul et al., 2023; Ansari et al., 2024; Woo et al., 2024; Das et al., 2024; Goswami et al., 2024), with work characterizing how forecasting quality scales with data and model size (Shi et al., 2024). Related work addresses missingness, irregular sampling, and longitudinal prediction in clinical time series (Shukla & Marlin, 2021; Che et al., 2018; Wornow et al., 2025; Fallahpour et al., 2024; Zhang et al., 2023). Wearable data share these challenges, with behavior-dependent observations spanning long personal histories.

Foundation models for wearables. Raw-biosignal models process 10- to 60-second PPG and ECG segments (Abbaspourazad et al., 2024; Pillai et al., 2025), while models of minute-resolution sensor features use five-hour to day-long windows and have scaled to 5.0M users (Narayanswamy et al., 2025; Xu et al., 2025; Narayanswamy et al., 2026). Models of higher-level behavior extend context to one week of hourly metrics from 162K users (Erturk et al., 2025). These approaches learn window-level representations for downstream prediction, with generative evaluations including imputation and minutes-to-hours forecasting (Narayanswamy et al., 2025; 2026). We instead use months of personal history for days-to-months physiological forecasting conditional on specified future behavior.

Breadth, depth, and context. Our corpus contains approximately 2B user-days: 1.8B observed user-days, or 2.2B calendar user-days including imputed intervening days (Table 1; Appendix A.1). The observed-day count alone is over 5× SensorFM's reported 350M sampled user-days. Relative to the wearable studies in Table 1, our maximum context spans approximately 52× as much history (365 days versus one week). Earlier work varies subject count and per-subject recording duration. Yuan et al. (2024) report downstream accuracy increasing with the log of pre-training subject count, with little change when per-subject recording time falls from twenty-four hours to six. In a different setting, Narayanswamy et al. (2025) find comparable performance when fewer subjects contribute more hours each at fixed total recording time. Both studies hold input windows fixed, so additional recordings supply more training windows rather than longer conditioning histories. Indeed, Narayanswamy et al. (2025) suggest that subject diversity may matter more as input windows grow from hours to days. We instead vary training history from 7 to 365 days and evaluate inference context separately, distinguishing the benefit of training on longer histories from that of providing more history to a trained model.

World models. World models learn representations of system state and dynamics that support control and planning (Ha & Schmidhuber, 2018; Hafner et al., 2025). Self-supervised video models

extend predictive representation learning to visual dynamics (Bardes et al., 2024), with work supporting action-conditioned prediction and planning (Assran et al., 2025). In health, HealthFormer (Lutsker et al., 2026) is a decoder-only model of physiological trajectories developed primarily on multi-visit clinical cohort data. We study action-conditioned physiological forecasting at wearable population scale, separating the roles of personal history, population size, and model capacity.

3 METHODS

3.1 TASK: ACTION-CONDITIONED FORECASTING

We forecast physiological trajectories conditional on a user's observed history and specified future actions and context. In forecasting evaluations, future action inputs are the user's observed actions over the forecast window. We evaluate daily forecast horizons of 1, 3, 7, 30, 90, and 180 days, spanning short-term deviations to seasonal changes.

We distinguish *training context*, the history length a model is trained to condition on, from *inference context*, the history supplied at evaluation. Inference context never exceeds training context. The main context sweeps use 7, 30, 90, 180, and 365 days, separating the benefits of longer histories during training from those of additional history at inference.

3.2 DATA

We use behavioral and physiological data from 5.47 million consumer users, spanning over 2 billion user-days. The data were collected in providing the service, and their research use is disclosed in the company's privacy notice. Direct identifiers were removed and replaced with pseudonymous keys, and access was restricted to authorized researchers. Input data to the models are daily summaries, one record per user per day. To our knowledge, this is the largest longitudinal wearable dataset used to train a health foundation model to date.

Daily records include cardiac, respiratory, temperature, blood oxygen, vascular, sleep, and activity measures; a hypnogram transition matrix; and sleep, readiness, activity, and illness scores. User-reported tags capture additional behavioral and health context, including sickness symptoms, alcohol and caffeine intake, medication use, travel, pregnancy, and menopause. We organize these variables into 17 semantic groups for tokenization and represent structured missingness with per-feature observation masks. Appendix A.2.1 lists variables, groups, availability, and missing-data handling.

The eight daily action covariates are sleep duration, deviation from usual sleep onset, active calorie expenditure, and tags for alcohol, GLP-1 medication use, late meals, travel, and caffeine. Forecast targets are 74 continuous sensor-derived biometric summaries. User-reported tags are inputs but not forecast targets; they are binary and sparsely observed.

Eligible users have at least 30 days of observed data. For pre-training and physiological forecasting, we partition participants into disjoint training, validation, and test sets. Validation and test each contain approximately 15,000 users, stratified to include 2,000 from each event-anchored cohort (Section 3.3). The remaining 5.44M users are available for pre-training, rounded to 5.4M.

3.3 COHORTS AND EVENT-ANCHORING

Population averages can conceal clinically relevant failure modes (Oakden-Rayner et al., 2020): forecasts may score well overall while failing to track illness recovery or changes following medication initiation. We therefore evaluate forecasting globally and in five event-anchored cohorts: alcohol, illness (mild), illness (severe), GLP-1, and postpartum. Evaluation windows are anchored to physiological or behavioral events, with horizons chosen for each event's timescale. Appendix A.2.2 gives cohort definitions.

3.4 MODEL

To study scaling, we use a common architecture family: a Transformer encoder with causal attention and an action-conditioned Transformer predictor. The encoder maps daily observations to per-day

latents. From a prefix of this latent stream, the predictor forecasts physiology conditional on actions supplied over the future window.

A two-layer gated recurrent network summarizes future actions into one vector per predicted step, concatenated with the encoder latent at the predictor input. A shared output head takes the latent state and a Fourier embedding of the query horizon, allowing one model to predict at different horizons. Training minimizes masked mean-squared error across target channels and forecast horizons. Appendix A.2.3 gives the horizon-sampling schedule and implementation details.

3.5 THE SCALING GRID

We study four axes. Three define a trained model, or *cell*: breadth (number of users), depth (days of per-user history), and capacity (five sizes from 1.3M to 303M parameters; Appendix Table 5). Inference context varies only at evaluation. Rather than train the full Cartesian product, we use controlled sweeps and matched comparisons: breadth ladders from 10K to 5.4M users at medium capacity and 100K to 5.4M at two larger capacities; isolated-depth sweeps from 7 to 365 days at three capacities; an eligibility-threshold sweep trading users for longer histories; and two sets matched on total user-days. Cells share an architecture family, optimizer, schedule, and effective batch size, and are trained to convergence. Appendix Table 6 lists all cells; Appendix A.2.5 gives hyperparameters and compute.

3.6 EVALUATION

Forecasting metric. We report R^2 on held-out targets. Our primary evaluation grid, labeled "marquee" in the figures, comprises 11 targets selected before scaling experiments: average HRV, lowest heart rate, readiness score, sleep score, deep sleep (minutes), temperature deviation, cardiovascular age gap, average SpO2, stressed duration, average respiration rate, and highest temperature. Across six cohorts and six horizons, this yields $\sim$231 populated panels after excluding inapplicable horizons (e.g., alcohol at $h > 1$). We summarize effects by their median and maximum across this grid and check consistency across all relevant targets, cohorts, and horizons unless otherwise stated. We also report other forecast targets where useful.

Autocorrelation baseline. The trailing-30-day-mean baseline predicts each target from its mean over the preceding 30 days. This captures persistence in wearable physiology; its R^2 provides a reference for task difficulty.

Representation probes. We fit linear probes on frozen representations to predict labels excluded from the pre-training objective, reporting R^2 for regression and AUROC for classification. *Subject-level* probes predict attributes such as chronic diagnoses and BMI, treated as constant per user. *Segment-level* probes predict day-specific labels such as cycle phase and self-reported sickness. For future labels, probes use the encoder state at t_{anchor} projected forward by h days through the trained action-conditioned predictor. Appendix A.2.8 describes the full pipeline.

Two-mechanism decomposition. We separate training-context and inference-context effects with matched comparisons. **Q1 (encoder quality)** compares 30-day-trained and 365-day-trained cells, both evaluated with 30 days of inference context, measuring the training-context effect at fixed inference history. **Q2 (user positioning)** compares 30 versus 365 days of inference context for the same 365-day-trained cell, changing only the supplied history. Both comparisons use the 1.7M-user M-model cells.

4 RESULTS

We first establish forecasting and transfer capabilities, then examine how they change with data, capacity, and context. We report forecasting R^2 on the primary target grid, disaggregated by cohort, and assess frozen representations on health prediction tasks.

4.1 FORECASTING AND TRANSFER ACROSS HEALTH TASKS

The model shows strong forecasting performance from days to months. The XL model, trained on 5.4M users with 365 days of context, well forecasts both raw signals and derived scores (Table 2).

Table 2: **Forecasting across horizons.** R^2 (gain over the trailing-30-day-mean baseline) for selected targets from XL trained on 5.4M users with 365-day context. Horizons $h \geq 30$ use smoothed targets: centered 7-day windows at h=30 and 15-day windows at h=90, 180 (Appendix A.2.3). The vertical rule separates raw from smoothed targets; scores are not directly comparable across it.

Target	h=1	3	7	30	90	180
Average HRV	.89 (+.10)	.85 (+.06)	.84 (+.06)	.91 (+.04)	.88 (+.08)	.85 (+.11)
Lowest heart rate	.90 (+.11)	.85 (+.08)	.84 (+.08)	.91 (+.06)	.87 (+.12)	.84 (+.16)
Sleep score	.70 (+.39)	.66 (+.35)	.65 (+.35)	.82 (+.19)	.78 (+.24)	.76 (+.29)
Readiness score	.72 (+.57)	.60 (+.49)	.55 (+.50)	.55 (+.57)	.51 (+.66)	.49 (+.68)
Temperature dev.	.37 (+.36)	.28 (+.31)	.26 (+.32)	.27 (+.49)	.17 (+.52)	.12 (+.47)
SpO_2 average	.46 (+.30)	.33 (+.17)	.27 (+.11)	.57 (+.10)	.47 (+.20)	.39 (+.27)

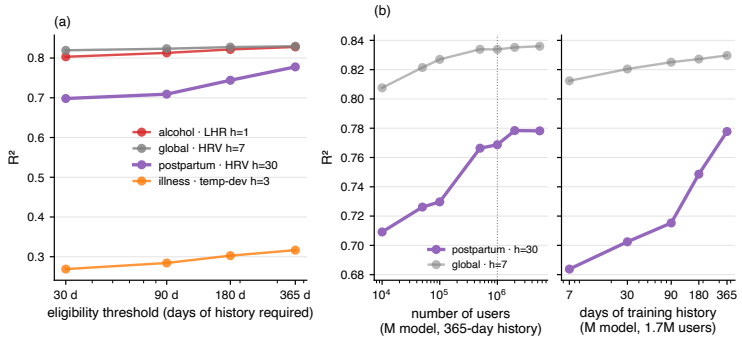

Figure 2: **Personal depth matters more than population breadth.** (a) Requiring longer histories improves forecasting despite reducing the eligible population. (b) At fixed capacity, breadth gains flatten near one million users (left), while depth gains continue through 365 days (right). The breadth and depth sweeps use different eligible populations; their points are not matched comparisons.

Performance remains strong in the challenging long-horizon postpartum setting: HRV forecasts reach $R^2 \approx 0.80$ at 90 days and 0.77 at 180 days. On the global cohort's primary grid, XL exceeds the trailing-30-day-mean baseline by a median of $+0.13$ R^2 per (target $\times$ horizon) panel, with larger gains on less autocorrelated targets such as composite scores and temperature. Performance and capacity dependence vary by target (Section 4.6); Appendix Figure 6 reports all 74 targets across cohorts and horizons.

Frozen representations transfer beyond the pre-training targets. Linear probes support both subject-level health prediction (Appendix Table 7) and day-specific event forecasting (Appendix Table 8). BMI regression reaches $R^2 = 0.76$; sleep apnea, type 2 diabetes, and hypertension probes reach AUROCs of 0.83, 0.89, and 0.84. Sickness forecasting reaches AUROC 0.83 three days ahead and 0.82 seven days ahead; next-day menstrual cycle stage reaches 0.95. None of these labels are pre-training targets. The same frozen representations therefore support both health assessment and anticipation of future events without task-specific backbone training.

4.2 Personal depth matters more than population breadth

Longer histories provide more value than additional users. At fixed users and capacity, extending training context from 7 to 365 days improves postpartum HRV forecasting by approximately $+0.09$ R^2, compared with $+0.02$ on the global population (Figure 2b). Trading users for history gives the same ordering: the 1.7M-user subset with 365 days outperforms 5.4M users with 30 days on every cohort, despite using 3.7M fewer users (Figure 2a).

Breadth saturates early; depth continues to help. With capacity and training context fixed, forecasting quality flattens near one million users, with little improvement from 1M to 5.4M (Figure 2b). Depth gains continue through 365 days. The 7-to-365-day lift is positive on nearly all cohort–target–horizon combinations across the capacities tested.

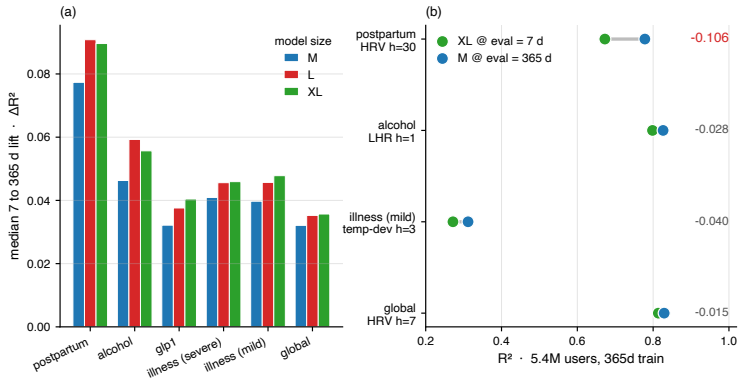

Figure 3: **Capacity increases the value of context.** (a) Median 7-to-365-day training-context gain by cohort and model size. (b) With 365 days of inference context, M outperforms XL given only seven days on the displayed panels.

Depth matters most for slower physiological changes. Postpartum recovery plays out over months, and its depth benefit grows at longer horizons; short perturbations such as respiratory illness gain little beyond a month of context. In this setting, the results favor accumulating longer histories over adding users with short histories. Section 4.5 separates the contributions of longer training and inference context.

4.3 Capacity increases the value of context

Larger models gain more from longer context. The 7-to-365-day lift is greater for L than M on 94% of primary panels, and for XL than L on 76%. The median lift is nevertheless flat from L to XL (Figure 3a), suggesting diminishing gains at the largest capacity. Larger models also outperform smaller models on forecasting R^2 and validation loss in most cells.

More capacity does not replace personal history. Given only seven days of inference context, XL underperforms M given 365 days on nearly every primary panel, by up to 0.11 R^2 on postpartum HRV (Figure 3b). The benefit of capacity therefore depends on the history available at inference, not just model size.

4.4 Scaling gains concentrate in event-anchored cohorts

Event-anchored evaluation reveals larger scaling gains. Median postpartum gains are approximately $2.4\times$ global gains for training context and $3.8\times$ for model size; breadth adds $+0.013$ R^2 versus near zero globally. Alcohol and illness gains are approximately 1.3–$1.4\times$ global gains for training context and 2.7–$3.1\times$ for model size (Appendix Table 11). The same contrast is visible on a single target: along the identical training-context axis, postpartum HRV performance improves almost five times as much as global HRV performance (Figure 2b).

The advantage over naive baselines also grows on event cohorts. On HRV panels, the trailing-30-day mean reaches $R^2 = 0.09$ for postpartum, 0.15 for alcohol, and 0.20 for severe illness, while the model remains near 0.7. These are settings where recent averages provide a weak forecast, yet the model retains substantial predictive performance. Across the primary grid, median gains over baseline are $+0.39$, $+0.34$, and $+0.29$ R^2, respectively—roughly two to three times the $+0.13$ global gain.

Population-only evaluation understates these benefits. Reporting only population-averaged R^2 would make scaling appear to offer limited improvement, despite substantially larger gains around events such as postpartum recovery and illness. Disaggregation therefore changes the interpretation of how much scaling helps, rather than simply adding detail to the global result. We recommend cohort-level reporting as standard practice for health foundation models, assessing improvements both globally and around physiological transitions and perturbations.

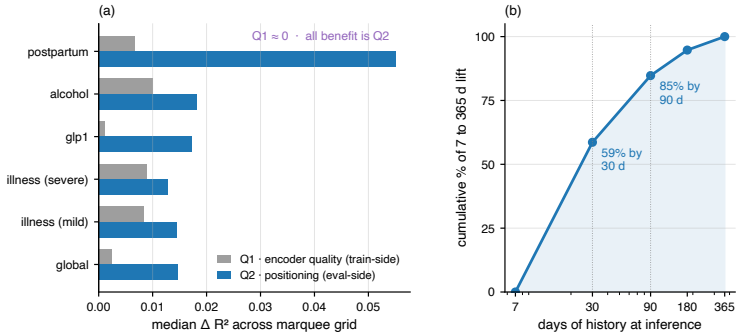

Figure 4: **Most of the context benefit comes at inference.** (a) Extending inference context (Q_2) contributes more than extending training context at fixed inference history (Q_1) in each cohort; postpartum Q_1 is near zero. (b) Most of the 7-to-365-day inference-context gain arrives within 30–90 days.

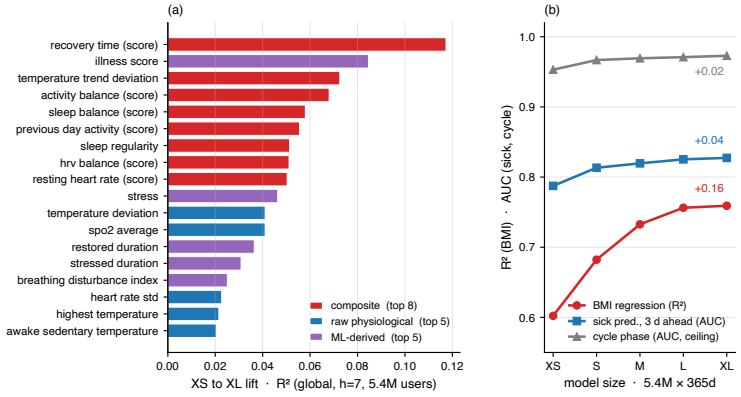

Figure 5: **Capacity gains depend on the target.** (a) XS-to-XL forecasting gains by target type. (b) BMI regression and sickness forecasting improve with capacity, while cycle-phase prediction remains near ceiling.

4.5 Most of the context benefit comes at inference

Using more history at inference contributes more than training on longer histories alone. The matched comparisons show a larger effect of extending inference context (Q_2) than training context at fixed inference history (Q_1) on every cohort (Figure 4). The difference ranges from approximately $+0.05$ R^2 for postpartum to $+0.004$ for severe illness. We report differences rather than ratios because Q_1 is near zero on several cohorts. For GLP-1 and postpartum, Q_1 is small relative to Q_2, with most of the measured context benefit coming from longer inference history.

Most of this benefit arrives within the first few months. Thirty days captures approximately 60% of the 7-to-365-day inference-context gain, and 90 days captures 85%. Comparisons that trade users for longer histories also favor inference context overall (Appendix Section A.3).

4.6 Capacity benefits composite targets more than raw signals

Composite targets benefit most from capacity. Increasing model size improves composite and derived scores by a median of $+0.04$ R^2, versus $+0.01$ for raw physiological targets (Figure 5). The recovery-time contributor has the largest composite gain, $+0.115$ R^2. HRV, heart rate, and respiratory rate change little beyond the smallest model. The contrast is sharper against the naive baseline: XL exceeds the trailing-30-day mean by a median of $+0.45$ R^2 on composite scores versus $+0.11$ on raw signals. Recent averages capture raw physiology more effectively than these multi-signal scores.

Transfer performance also scales differently across targets. BMI R^2 increases from approximately 0.60 at XS to 0.76 at larger sizes, while three-day sickness AUROC rises from 0.79 to 0.83. XL performs best on most other probes; small label counts make capacity trends harder to interpret for stroke and coronary heart disease (Appendix Table 7). The apparent benefit of model scaling therefore depends on the targets evaluated; we recommend reporting gains by target type.

5 DISCUSSION

This work advances a shift from wearable models that summarize observed signals toward models that represent and predict physiological change. The scaling results suggest that this shift depends not only on larger populations or models, but on preserving longitudinal context. Within the ranges studied, longer personal histories improve forecasting more than additional users, with much of the benefit coming from inference context. A practical implication is that model capacity should be chosen alongside the personal history available at deployment, rather than scaled in isolation.

Evaluation design also changes what scaling appears to achieve. Larger gains on event-anchored cohorts and composite targets show why population-wide averages are insufficient to characterize model improvements. Progress should therefore be assessed across physiological targets, forecast horizons, and periods of physiological change, not only across users in aggregate. These differences also caution against treating a single scaling curve as a complete description of model capability.

The learned representations support a broader set of predictions than the biometric forecasting objective alone. Frozen representations support subject-level health prediction, while forward-projected representations support forecasts of self-reported sickness and menstrual-cycle phase. These results suggest that learning physiological dynamics can provide a shared representation for multiple health prediction tasks, without learning a new representation for each downstream label.

Our study is limited to one wearable platform and a population that skews young, female, and United States based. Cross-device and cross-population validation remain necessary; breadth saturation here does not imply that population diversity is unimportant. We also treat the pre-training corpus as fixed and uniformly weighted, leaving the effects of cohort and event sampling unresolved. Several downstream tasks use self-reported labels, motivating further validation against independently verified outcomes. Prospective evaluation should also distinguish forecasts conditioned on specified future behavior from those made when that behavior is unknown.

Action conditioning offers a further direction: comparing predicted trajectories under different behavioral scenarios. The architecture supports such comparisons, but observational forecasting accuracy does not establish that they recover the effects of changing behavior. Progress toward intervention-oriented models will require addressing confounding and validating the assumptions underlying causal interpretation. Whether these models can reliably guide behavioral decisions remains an open question.

The broader goal is a model of health that develops with an individual's history: incorporating new observations, tracking physiological changes, and anticipating how those changes may unfold. Rather than interpreting sleep, activity, and cardiovascular measurements separately, such a model could relate them to a person's evolving health trajectory. This work provides a foundation for that direction by showing how to scale long-history forecasting and learn representations that support multiple health prediction tasks. Wearable world models could help people understand not only their current health, but where it may be heading and how their choices might change that course.

Reproducibility statement The data used in this work contain sensitive health information, and we cannot release the raw data or the trained model weights. We instead document the components needed to reproduce our findings on comparable data. Appendix Table 6 gives the full training configuration for every experiment cell, including data volume along each scaling axis, while Appendix Table 5 provides details on model architecture and parameter count. Appendix A.2.1 describes signal derivation, temporal aggregation, and cohort construction, including inclusion criteria and handling of missing data. Appendices A.2.5 and A.2.8 describe the training protocol and each downstream task: label definition and provenance, cohort, evaluation split, and metric.

Ethics statement This study used pre-existing consumer data, pseudonymized as described in Section 3.2. This dataset is not available for download. Data requests are handled through the company's research-collaboration review process. We are not making our code available. No analysis used direct identifiers. Longitudinal physiological records may carry re-identification risk even without direct identifiers, so we do not release raw data or model weights. Our cohort consists of people who own and consistently wear a consumer smart wearable device, who differ from the general population in age, income, health engagement, and geography. Reported performance may not transfer to populations with less access to these devices, and progress in this area risks widening existing gaps in who benefits from health technology. Realizing the benefits of models like ours will require validation in more representative cohorts and attention to who can access the resulting tools.

AI use statement In this work, we used generative AI tools for helping provide critiques and feedback on research methodology, assist in implementing an experimental pipeline, and assist in data analysis. Additionally, we used generative AI tools for providing feedback on literature searches, writing, and helping iterate on new figures and tables. We have reviewed all AI-assisted work. LLM-generated text, figures, tables, and results were vetted for correctness and edited when necessary rather than taken verbatim. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

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)*, 2024. arXiv:2312.05409.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *Transactions on Machine Learning Research (TMLR)*, 2024. arXiv:2403.07815.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research (TMLR)*, 2024. V-JEPA; arXiv:2404.08471.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1):6085, 2018.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2310.10688.

Eray Erturk, Fahad Kamran, Salar Abbaspourazad, Sean Jewell, Harsh Sharma, Yujie Li, Sinead Williamson, Nicholas J. Foti, and Joseph Futoma. Beyond sensor data: Foundation models of behavioral data from wearables improve health predictions. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2507.00191.

Adibvafa Fallahpour, Mahshid Alinoori, Wenqian Ye, Xu Cao, Arash Afkanpour, and Amrit Krishnan. Ehrmamba: Towards generalizable and scalable foundation models for electronic health records. *arXiv preprint arXiv:2405.14567*, 2024.

Azul Garza, Cristian Challu, and Max Mergenthaler-Canseco. Timegpt-1. *arXiv preprint arXiv:2310.03589*, 2023.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2402.03885.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2203.15556.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, 2021.

Guy Lutsker, Gal Sapir, Jordi Merino, Smadar Shilo, Anastasia Godneva, Eli Meirom, Shie Mannor, Hagai Rossman, Gal Chechik, and Eran Segal. Simulating clinical interventions with a generative multimodal model of human physiology. *arXiv preprint arXiv:2604.27899*, 2026.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.16264.

Girish Narayanswamy, Xin Liu, Kumar Ayush, Yuzhe Yang, Xuhai Xu, Shun Liao, Jake Garrison, Shyam Tailor, Jacob Sunshine, Yun Liu, et al. Scaling wearable foundation models. In *International Conference on Learning Representations*, volume 2025, pp. 59382–59413, 2025.

Girish Narayanswamy, Maxwell A Xu, A Ali Heydari, Samy Abdel-Ghaffar, Marius Guerard, Kara Vaillancourt, Zhihan Zhang, Jake Garrison, Levi Albuquerque, Dimitris Spathis, et al. Towards a general intelligence and interface for wearable health data. *arXiv preprint arXiv:2605.22759*, 2026.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *ACM Conference on Health, Inference, and Learning (CHIL)*, 2020.

Arvind Pillai, Dimitris Spathis, Fahim Kawsar, and Mohammad Malekzadeh. Papagei: Open foundation models for optical physiological signals. In *International Conference on Learning Representations*, volume 2025, pp. 48230–48261, 2025.

Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Arian Khorasani, George Adamopoulos, Rishika Bhagwatkar, Marin Biloš, Hena Ghonia, Nadhir Vincent Hassen, Anderson Schneider, et al. Lag-llama: Towards foundation models for time series forecasting. *arXiv preprint arXiv:2310.08278*, 2023.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

Jingzhe Shi, Qinwei Ma, Huan Ma, and Lei Li. Scaling law for time series forecasting. *Advances in Neural Information Processing Systems*, 37:83314–83344, 2024.

Satya Narayan Shukla and Benjamin M. Marlin. Multi-time attention networks for irregularly sampled time series. In *International Conference on Learning Representations (ICLR)*, 2021.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2402.02592.

Michael Wornow, Suhana Bedi, Miguel Angel Fuentes Hernandez, Ethan Steinberg, Jason Fries, Christopher Ré, Sanmi Koyejo, and Nigam Shah. Context clues: evaluating long context models for clinical prediction tasks on ehr data. In *International Conference on Learning Representations*, volume 2025, pp. 16518–16555, 2025.

Yu Yvonne Wu, Yuwei Zhang, Hyungjun Yoon, Ting Dang, Dimitris Spathis, Tong Xia, Qiang Yang, Jing Han, Dong Ma, Sung-Ju Lee, et al. Wearable foundation models should go beyond static encoders. *arXiv preprint arXiv:2603.19564*, 2026.

Maxwell A Xu, Girish Narayanswamy, Kumar Ayush, Dimitris Spathis, Shun Liao, Shyam A Tailor, Ahmed Metwally, A Ali Heydari, Yuwei Zhang, Jake Garrison, et al. Lsm-2: Learning from incomplete wearable sensor data. *arXiv preprint arXiv:2506.05321*, 2025.

Hang Yuan, Shing Chan, Andrew P Creagh, Catherine Tong, Aidan Acquah, David A Clifton, and Aiden Doherty. Self-supervised learning for human activity recognition using 700,000 person-days of wearable data. *NPJ digital medicine*, 7(1):91, 2024.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. arXiv:2106.04560.

Jiawen Zhang, Shun Zheng, Wei Cao, Jiang Bian, and Jia Li. Warpformer: A multi-scale modeling approach for irregular clinical time series. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2023.

A Appendix

A.1 Additional Discussion of Scale of Pre-Training Corpus

Table 1 details. Corpora are reported in different units. Calendar coverage, recorded sensor-hours, and sampled window-hours measure different aspects of a corpus; differences reflect recording schedules, sampling, and missingness. Where only recorded hours are reported, we divide by 24 to obtain duration equivalents rather than distinct calendar user-days. [†]For the waveform models, calendar coverage overstates data volume by orders of magnitude, because the corpora are sparse short segments drawn from a long span: the Apple PPG corpus spans 93 days per participant on average but contains $19.9\text{M} \times 60\,\text{s} \approx 0.33\text{M}$ hours of signal in total ($20.8\text{M} \times 10\,\text{s} \approx 58\text{K}$ hours for PaPaGei). [‡]LSM-1 samples ten five-hour windows per subject, so its 40M hours are window-hours and calendar coverage is not well defined; we report 24-hour equivalents. That corpus is also subject-imbalanced, and the balanced corpus used for most of its scaling results is 6.6M hours (0.3M day-equivalents). For SensorFM we report the 350M human-days sampled (10 weeks from each of 5.0M subjects), corresponding to over 2B hours of retained multimodal sensor data; the 175M "day-long samples" reported for its largest pre-training volume counts overlapping sliding windows (stride randomized over 8–24 minutes) rather than distinct days. LSM-2 reports both figures directly: $3{,}581{,}748$ person-days and 40M hours, an average of 11.2 recorded hours per covered day. Our corpus contains 1.8B observed user-days and 2.2B calendar user-days when imputed days between observations are included. We describe this scale as approximately 2B user-days in the main text. The imputed days do not constitute additional observed data. Using the observed-day count alone, our corpus contains over $5\times$ SensorFM's reported 350M sampled human-days. Our own eligibility threshold couples breadth to depth: 5.47M users have at least 30 days of history, falling to 4M at 90 days, 3M at 180 days, and 1.7M at 365 days. Table 1 reports the 5.44M pre-training users drawn from the eligible population, rounded to 5.4M in the main text. We state the history threshold wherever a per-user recording span is quoted.

A.2 Additional Methods

A.2.1 Dataset composition and missingness

Table 3 reports demographics and self-reported health history across the train, validation, and test splits of the 5,473,776 users in the corpus. The population skews young (median age 36) and female

Table 3: Demographic and health-history distribution of the study population across train/val/test splits. Continuous fields report mean ± std and median [Q1, Q3]; binary/categorical fields report N (% of the split). Diagnoses and blood-pressure measurements come from a self-reported cardiovascular health questionnaire and are shown as % of users who completed it.

Field	Train	Val	Test
N users	5,443,781	14,999	14,996
Age (years), mean ± std	38.7 ± 13.2	38.1 ± 12.0	38.1 ± 12.0
Age (years), median [Q1, Q3]	36.0 [28.0, 47.0]	35.0 [29.0, 46.0]	35.0 [29.0, 46.0]
missing	0.2%	0.1%	0.1%
BMI, mean ± std	26.3 ± 5.6	25.8 ± 5.3	25.8 ± 5.2
BMI, median [Q1, Q3]	25.2 [22.4, 29.0]	24.7 [22.2, 28.3]	24.8 [22.2, 28.2]
missing	0.3%	0.1%	0.2%
Male, N (%)	1,697,046 (31.2%)	3,329 (22.2%)	3,392 (22.6%)
Female, N (%)	3,746,735 (68.8%)	11,670 (77.8%)	11,604 (77.4%)
Sex missing, N (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Top country, name (%)	USA (74.8%)	USA (70.1%)	USA (70.3%)
Country missing, N (%)	0.3%	0.2%	0.1%
Completed CV questionnaire, N (%)	445,382 (8.2%)	3,570 (23.8%)	3,600 (24.0%)

(68.8%), and is majority United States based. Validation and test are more heavily female than train (roughly 78% versus 69%) by construction, as held-out splits are stratified to guarantee coverage of the event-anchored cohorts of Section 3.3, including postpartum.

Table 4: Per-feature coverage and missingness across the study population (N = 5,473,776 users). % Users is the fraction of users with $\geq$ 1 non-null day for the feature. The 16 hypnogram transition probabilities are collapsed into a single row as they are captured together. The final three channel groups relate to user-tagged events corresponding to 1) illness events, 2) women's health events, and 3) procedures and medications, where we are unable to report specific tag values.

Domain	Group	Feature	% Users	Avg. Missing	Global Null Frac
Nighttime cardio	Heart & HRV	Avg HRV (sleep)	99.97%	20.01%	22.82%
		HRV std (sleep)	99.97%	20.02%	22.83%
		Min HR (sleep)	99.98%	19.99%	22.79%
		Avg HR (sleep)	99.99%	19.93%	22.72%
		HR std (sleep)	99.97%	20.02%	22.83%
Respiratory	Respiratory	Avg respiration rate	99.99%	19.82%	22.64%
		Respiration variability	99.99%	19.82%	22.64%
Temperature	Temperature	Max skin temp (sleep)	99.99%	19.92%	22.71%
		Skin temp deviation	99.99%	20.02%	22.81%
		Skin temp trend deviation	99.99%	23.11%	25.59%
		Awake sedentary temp	98.87%	22.85%	28.79%
Respiratory	SpO_2	Avg SpO_2	99.60%	26.86%	29.96%
		Breathing disturbance index	99.71%	23.62%	28.52%
Vascular age	Vascular	Age − vascular age	97.42%	28.79%	37.26%
		Vascular age	97.43%	28.79%	37.25%
		Vascular age variance	97.43%	28.79%	37.25%
Sleep	Sleep	Deep sleep (min)	100.00%	19.92%	22.71%
		REM sleep (min)	100.00%	19.92%	22.71%
		Light sleep (min)	100.00%	19.92%	22.71%
		Awake during sleep (min)	100.00%	19.92%	22.71%
		Sleep latency	99.43%	21.64%	26.59%
		Sleep regularity	99.75%	24.94%	27.08%
		Time offset of min HR	99.98%	19.98%	22.78%
		Wake-time offset (h)	99.99%	19.53%	22.18%
		Sleep score	99.99%	19.92%	22.71%

Continued on next page

Domain	Group	Feature	% Users	Avg. Missing	Global Null Frac
		Readiness score	99.99%	19.98%	22.77%
	Hypnogram	Sleep-stage transition probs (16)	99.86%	19.94%	22.83%
Readiness	Readiness sub-scores	Resting HR sub-score	99.98%	20.06%	22.85%
		Body temp sub-score	99.99%	20.00%	22.78%
		Recovery index sub-score	99.99%	20.00%	22.80%
		Activity balance sub-score	99.98%	21.07%	24.02%
		Recovery time sub-score	99.97%	20.63%	23.45%
		HRV balance sub-score	99.97%	25.02%	26.39%
		Sleep balance sub-score	99.99%	23.62%	25.64%
		Restfulness sub-score	99.99%	19.92%	22.71%
		Prior-day activity sub-score	99.87%	24.13%	27.62%
Stress	Stress	Stress score	97.90%	36.09%	41.33%
		Time stressed (min)	98.81%	22.84%	28.83%
		Time recovered (min)	98.81%	22.84%	28.83%
Activity	Activity	Steps	99.97%	20.63%	23.45%
		Activity score	99.97%	20.63%	23.45%
		Sedentary time	99.97%	20.63%	23.45%
		Resting time	99.97%	20.63%	23.45%
		Low-activity time	99.97%	20.63%	23.45%
		Medium-activity time	99.97%	20.63%	23.45%
		High-activity time	99.97%	20.63%	23.45%
		Non-wear time	99.97%	20.63%	23.45%
		Inactivity alerts	99.97%	20.63%	23.45%
Daytime cardio	Awake HR	Avg awake HR	99.85%	19.78%	22.63%
		Max awake HR	99.85%	19.78%	22.63%
		Min awake HR	99.85%	19.78%	22.63%
		Median awake HR	99.85%	19.78%	22.63%
		Restorative awake time	99.85%	19.78%	22.63%
		Awake sedentary min HR	98.87%	22.85%	28.79%
		Awake sedentary HRV	98.87%	22.85%	28.79%
Fitness	VO_2max	VO_2max	47.84%	85.86%	95.27%
		VO_2max confidence	17.89%	77.01%	97.64%
		External VO_2max reading count	38.54%	92.60%	97.51%
Illness	Illness	Illness detection score	94.94%	32.54%	46.15%
Metabolic	CGM	Avg blood glucose	1.95%	87.30%	99.80%
		Max blood glucose	1.95%	87.30%	99.80%
		Median blood glucose	1.95%	87.30%	99.80%
		Min blood glucose	1.95%	87.30%	99.80%
		Time glucose above range	2.03%	87.50%	99.80%
		Waking glucose	1.19%	87.58%	99.91%

We include information about the variables used, their missingness, and the group assignments in Table 4. The data have substantial and structured missingness. Some measurements arrive only during sleep, some only during periods of activity, and others are opportunistic and depend on how the user interacts with the device. Most features are observed on nearly every user and on roughly 80% of that user's days, including marquee forecast targets used in the main paper. Coverage is small for a few opportunistic groups: VO_2max, which was introduced to the wearable platform partway through the study, and continuous glucose. We do not impute missing observations with a single global value. We use last one carried forward observation for biometrics, and zero-imputation for events and activities for when last one carried forward would not make intuitive sense. Every raw feature is accompanied by a smoothed version over a short window and a binary mask that records whether the raw value was observed on that day, so the model can see missingness explicitly and learn to weight imputed values accordingly. This prevents the model from being dominated by imputed zeros for glucose while still allowing it to consume the signal on the days it is available.

A.2.2 Details on Cohort Definitions

The alcohol cohort anchors on any day that carries a user-logged alcohol tag and is evaluated at a one-day horizon, capturing next-day effects on sleep and cardiovascular metrics. Mild illness anchors on the first day of a sickness, cold, or fever tag within users who have a history of logging mild illness reliably, and is evaluated at horizons of one, three, and seven days which cover the acute phase and the tail of recovery. Severe illness uses the same first-day anchoring but restricts to influenza and COVID-19 tags within the corresponding severe-illness cohort. GLP-1 anchors on the first day of a GLP-1 medication tag, and is evaluated at horizons from one week through six months, which span the slow-rolling physiological changes that these medications induce. Postpartum uses the same long-horizon evaluation and anchors on the first day of the postpartum interval. For GLP-1 and postpartum, which use longer-term evaluation, we require at least 90 days of prior history from the anchor event to establish a pre-event baseline.

Consecutive event anchors within a single user are spaced at least seven days apart to prevent overlap between adjacent evaluation windows, and we cap the number of anchors any single user contributes to a cohort at between three and ten, depending on how common the event is. This prevents any one heavy-tag user from dominating a cohort. A cohort's effective sample size is set by the number of events instead of by the number of users.

A.2.3 Additional Model Details

Tokenization. We tokenize the daily grid at the day level, one token per semantic group per day. Each of the 17 groups defined in Section 3.2, as well as the group of actions, has its own linear projection whose input width matches the number of channels in that group, so all group tokens share a common latent width while being computed independently. Keeping groups on separate projections gives sparse modalities their own parameter budget and prevents them from being averaged away by the dense ones. Actions enter the encoder as one additional channel group, so past actions are visible alongside physiology in the causal encoder without mixing into the projections that tokenize the biometric channels.

Encoder Details. Both the encoder backbone and the predictor use rotary position embeddings and apply QK-normalization inside self-attention. Feed-forward blocks follow the standard Transformer template with a GELU activation and hidden width four times the model dimension. Continuous glucose readings are present in fewer than one in a five hundred user-days, so we exclude that group from the encoder's attention on days when it is absent, similar to approaches in past work when large amounts of missingness are present (Xu et al., 2025). We keep missing values, along with a missingness mask, in the encoder's attention for the other variables, as missingness for those may encode important information that may be lost if removed.

Objective. We train with a masked mean-squared error over all target channels and forecast horizons. On any position where the target is missing due to a sensor gap or falls beyond the end of a user's observed data, the loss is masked out and the model receives no gradient at that position. The horizon losses are equally weighted.

Long-horizon targets are lightly smoothed to remove day-to-day noise that no model can reasonably predict. Horizons up to and including one week are trained on raw next-day values. Horizons between 15 and 60 days use a centered 7-day window on the target, and any horizons larger than 60 days use a centered 15-day window. This lets us report long-horizon quality without penalizing the model for missing a single noisy day.

Optimization. We train with AdamW using beta parameters 0.9 and 0.97, weight decay 0.05, and no weight decay on biases. Peak learning rate follows an inverse-square-root scaling rule in model width, giving the values reported in Table 5. The learning rate warms up linearly over the first 5% of optimizer steps and follows a cosine decay to one-tenth of the peak over the remainder. Effective batch size is fixed at 256 examples across every cell in the grid, achieved by tuning per-device batch size, gradient accumulation, and data-parallel replicas together. Training is done in bfloat16 mixed precision.

Table 5: Model configurations at each step of the capacity ladder. Width, depth, and predictor depth scale together across the five sizes. Head dimension is held at 64 throughout, so the number of attention heads grows with width. Feed-forward width is four times the model width. Peak learning rate is set inverse to the square root of model width, anchored at 10^{-3} for XS.

Size	Width	Encoder layers	Predictor layers	Heads	FFN width	Peak LR	Params (M)
XS	128	4	2	2	512	1.0×10^{-3}	1.3
S	256	8	4	4	1024	7.0×10^{-4}	9.7
M	384	12	6	6	1536	6.0×10^{-4}	32.3
L	768	12	6	12	3072	4.0×10^{-4}	128.3
XL	1024	16	8	16	4096	3.5×10^{-4}	303.3

A.2.4 Additional Scaling Grid Details

Capacity sweeps. We train five model sizes, XS, S, M, L, and XL, which scale some combination of width, depth, and predictor depth together. The five sizes span roughly 1.3 to 300 million trainable parameters, and full configurations are given in the Appendix (Table 5). Unless stated otherwise, a step along the capacity ladder means moving from one row of this table to the next. We train the full capacity ladder at multiple anchor points across the other axes described below.

Breadth ladder. Holding depth at 365 days, we sweep breadth from ten thousand to over 5 million users. The full sweep is run at model size M over seven breadth points: 10K, 50K, 100k, 500K, 1M, 2M, and 5.4M users. We repeat the sweep at L and XL over five breadth points from 100K to 5M users, so that the breadth-scaling curve can be measured at three sizes and any interaction between breadth and capacity is visible in the same experiment. A parallel M-size ladder at 90-day depth spans 10K to 5M users, letting us check whether the scaling behavior differs when each user contributes less history.

Isolated depth. Holding breadth fixed at 1.7 million users and varying depth over 7, 30, 90, 180, and 365 days at three model sizes (M, L, and XL), we isolate the effect of per-user history and how it interacts with capacity. This is the main axis for the training-context story, since only the amount of history per user changes.

Breadth-for-depth trade. In practice, requiring more history per user shrinks the eligible population. We follow this trade explicitly by varying an eligibility threshold on required per-user history, which couples breadth to depth. Raising the threshold from 30 to 365 days walks the eligible population from 5.4M users with 30 days each, to 4M users with 90 days, to 3M users with 180 days, to 1.7M users with 365 days. The longer-history populations are nested subsets of the shorter-history ones, so this experiment measures what the eligibility filter buys in depth against the users it removes.

Capacity and inference context at scale. At the maximum breadth and depth (5.4 million users, 365-day training context), we evaluate model sizes M, L, and XL across the full inference-context sweep of 7, 30, 90, 180, and 365 days. These are the flagship cells for the capacity analysis.

Matched user-days. To compare breadth against depth at equal data volume, we hold total user-days fixed and contrast broad-but-shallow cells against narrow-but-deep cells at two budgets: 164 million user-days and 510 million user-days. These matched sets are the basis for the breadth-versus-depth attribution described in Section 3.6. The full specification of these experiments can be found in Appendix A.2.7.

Further training details, including hyperparameter choices and compute budget, can be found in Appendix A.2.5.

A.2.5 Training protocol

Hyperparameters. Beyond the width, depth, and learning-rate values reported in Table 5, cells differ only in per-device batch size and gradient accumulation, which are tuned per cell to hold the effective batch size at 256 while fitting the model in an 80GB accelerator's memory. Table 6 lists the breadth, depth, and capacity setting of every cell..

Hardware and compute. Training runs on A100-80GB accelerators, with one to four data-parallel replicas per cell chosen to keep wall-clock reasonable at the largest breadth and capacity settings.

A.2.6 Additional Evaluation Details

Anchor floor. We consider the same evaluation floor for all windows, which requires at least 30 days of prior in-window history. This ensures every cell in the grid is compared on the same set of days and targets, so differences in performance across cells reflect the model, not shifts in the evaluation population.

Target taxonomy. We label each target as **raw** if it is a direct sensor-derived measurement (HRV, HR, temperature), **composite score** if it is a derived score or a contributor to it, or **ML-derived** scores if it is derived using machine-learning algorithms from raw measurements (stress information, sleep-staging information). We break scaling effects out by this taxonomy because the composite scores are algorithmic aggregations of many raw signals and thus expose different requirements of the model.

Breadth-versus-depth attribution. Using the matched-user-day cells, we compare the broad and deep cells at strict-overlap inference context to measure the encoder-quality component of depth's advantage, and at each cell's longest supported inference context to measure the component that comes from consuming longer history at inference.

Subgroup evaluation. To test whether the findings hold across demographic subpopulations, we re-slice the evaluation of already-trained cells by age band (18–29, 30–44, 45–59, 60+) and self-reported biological sex (Koh et al., 2021; Sagawa et al., 2020). No additional training is required. See details in Appendix A.3.1.

A.2.7 Matched User-Days Experiments

At roughly 164 million user-days we compare 5.4M users at 30 days each, 2.3M users at 70 days, and 450K users at 365 days. At roughly 510 million user-days we compare 3.9M users at 130 days, 1.7M users at 300 days, and 1.13M users at 450 days. Within each budget, cells are matched on total user-days.

A.2.8 Training Probes

We consider two families of probes for predicting held-out labels.

Subject-level probes predict time-invariant properties of the user (BMI and selected clinical diagnoses). For each user, encoder embeddings from multiple sampled anchor windows are averaged to obtain a single **subject embedding**, and the probe is fit at the subject level, similar to past work Erturk et al. (2025); Xu et al. (2025); Narayanswamy et al. (2026). Because labeled clinical outcomes are scarce, the results in the main paper pool the train, validation, and test splits and then take a 70/30 probe train/test split. Our models are never trained with knowledge of these outcomes, so pooling the splits leaks no information the model learned during pre-training. For completeness, Table 10 reports the same targets under the standard validation-fit, test-scored protocol; those numbers rest on far fewer users and lower positive rates, and we read them as a floor rather than as our estimate.

Segment-level probes predict labels attached to a specific day t_{anchor}: cycle phase and sickness. As mentioned in the main paper, for labels at future horizons ($h > 0$), the probe input is the encoder state at t_{anchor} rolled forward by h days through the trained action-conditioned predictor. This ensures the probe scores a representation of the future that the model itself would have generated. For binary labels with prevalence imbalance (e.g., sickness), we report AUROC on a matched case-control subset (positive anchor days paired with matched negative days from the same user cohort) so that the score reflects pre-event physiology. Probes are fit on the held-out validation set and evaluated on the test set such that no user is shared between both.

A.3 Additional Results

The mechanism holds when more depth requires fewer users. Figure 4 varies per-user history at a fixed population. The trade a practitioner often faces is that using longer histories shrinks the

eligible training population. Repeating the comparison at equal total user-days, at two budgets, gives the same answer. Positioning exceeds encoder quality on 4 of 6 cohorts at both budgets, with postpartum showing by far the largest gap ($Q_2 = +0.067$ at 164M user-days and $+0.081$ at 510M). The two illness cohorts are the exception, and their ordering here is the reverse of Figure 4. No illness cell in either experiment exceeds 0.011 R^2 in absolute value, and hence, we read this as evidence that neither channel is active for these cohorts. Depth still improves forecasting performance in the context of illness recovery by more than it improves for the global population (Table 11). Our result shows that this decomposition cannot resolve where that improvement comes from. Depth primarily helps by allowing a model to use more context, and this finding survives shrinking the eligible population to obtain that depth.

A.3.1 Findings reproduce across age and sex subgroups

To test whether the scaling behaviors in the main paper reflect a majority subgroup or a robust property of the model, we re-slice the test-split evaluation of the already-trained cells into eight subgroups formed by four age bands (18-29, 30-44, 45-59, 60+) crossed with self-reported biological sex. The eight subgroups range in size from 392 users (60+, male) to 5,371 users (30-44, female). We compute R^2 per subgroup against the subgroup-specific target mean. We focus results on the global cohort. We aggregate across the same 11-target marquee grid used throughout Section 4. We reproduce each of the figures in the main paper for each subgroup.

Depth remains the important axis for scaling. The $7 \rightarrow 365$-day training-context lift is positive on $\geq 98\%$ of (target $\times$ horizon) panels within every subgroup, with median lifts of $+0.030$ to $+0.039$ R^2. Absolute gains vary modestly across subgroups within this tight band without a clean age trend. Breadth likewise saturates near one million users in every subgroup, and trading users for per-user history stays net-positive in every subgroup.

Capacity and context stay complementary. Figure 8 reproduces both claims from Section 4.3 within subgroups. The L model shows a steeper $7 \rightarrow 365$-day lift than the M model on 61% to 77% of marquee panels within every subgroup. XL shows a steeper lift than L on 44% to 79% of panels. The endpoint ordering XL compared to M holds on 91% to 99% of marquee panels within every subgroup. However, with only seven days of inference context, the XL model loses to the M model with a full year of inference context across all demographic subgroups, with negative gaps ranging from -0.025 (60+, female) to -0.043 (30-44, female) R^2. The slight discrepancy between the subgroup results and global results could be a result of smaller subgroup panel counts, resulting in noisier medians.

Positioning dominates encoder quality per subgroup, with most benefits arriving at early evaluation context. Figure 9a decomposes the context benefit across all demographic subgroups. The evaluation-side positioning contribution (Q_2) exceeds the train-side encoder-quality contribution (Q_1) on every one of the eight subgroups, by factors of 3.6$\times$ to 8.3$\times$. This is most apparent for females between ages 30-44 (8.3$\times$) and 18-29 (7.2$\times$), both showing a similar pattern and likely, in part, due to postpartum effects. In these subgroups, encoder quality contributes minimally and nearly all of the context benefit is attributable to inference context.

The cumulative evaluation-context saturation (Figure 9b) tracks the population curve in every subgroup. All subgroups capture 55% to 62% of the total lift by 30 days of inference context, 79% to 87% by 90 days, and 92% to 96% by 180 days. The fastest-saturating subgroup is 18-29 female (62% by 30 days); the slowest is 60+ female (55% by 30 days, 92% by 180 days).

Where subgroup performance differs. Absolute R^2 is systematically lower for older users (60+ female HRV h=7 at 5.4M $\times$ 365d: ≈ 0.71; 18-29 female same cell: ≈ 0.83). We hypothesize that the difference reflects lower physiological variance in older users rather than model failure. The scaling curves themselves, meaning how much each additional day of history, each additional million users, each additional order of magnitude of capacity, and each additional month of inference context buys the model, have the same shape in every subgroup we test. Performance is strong in every subgroup, including those least represented in the population. Closing the remaining gap for older users is future work.

Table 6: Trained cells in the scaling grid. Each row is one trained model. Cells that appear in multiple sub-experiments (e.g. M/L/XL cells at 100K×365d and 5.4M×365d, which serve both the breadth ladder and the capacity sweep; M @ 1.7M×365d, which serves both isolated depth and breadth-for-depth) are listed once under their primary sub-experiment.

Sub-experiment	Size	Breadth	Depth (days)
Breadth ladder	M	10K	365
Breadth ladder	M	50K	365
Breadth ladder	M	100K	365
Breadth ladder	M	500K	365
Breadth ladder	M	1M	365
Breadth ladder	M	2M	365
Breadth ladder	M	5.4M	365
Breadth ladder	L	100K	365
Breadth ladder	L	500K	365
Breadth ladder	L	1M	365
Breadth ladder	L	2M	365
Breadth ladder	L	5.4M	365
Breadth ladder	XL	100K	365
Breadth ladder	XL	500K	365
Breadth ladder	XL	1M	365
Breadth ladder	XL	2M	365
Breadth ladder	XL	5.4M	365
Breadth ladder (90d)	M	10K	90
Breadth ladder (90d)	M	50K	90
Breadth ladder (90d)	M	100K	90
Breadth ladder (90d)	M	500K	90
Breadth ladder (90d)	M	1M	90
Breadth ladder (90d)	M	2M	90
Breadth ladder (90d)	M	5.4M	90
Capacity sweep	XS	100K	365
Capacity sweep	S	100K	365
Capacity sweep	XS	5.4M	365
Capacity sweep	S	5.4M	365
Isolated depth	M	1.7M	7
Isolated depth	M	1.7M	30
Isolated depth	M	1.7M	90
Isolated depth	M	1.7M	180
Isolated depth	M	1.7M	365
Isolated depth	L	1.7M	7
Isolated depth	L	1.7M	30
Isolated depth	L	1.7M	90
Isolated depth	L	1.7M	180
Isolated depth	L	1.7M	365
Isolated depth	XL	1.7M	7
Isolated depth	XL	1.7M	30
Isolated depth	XL	1.7M	90
Isolated depth	XL	1.7M	180
Isolated depth	XL	1.7M	365
Breadth-for-depth trade	M	5.4M	30
Breadth-for-depth trade	M	4M	90
Breadth-for-depth trade	M	3M	180
Matched user-days (164M)	M	5.4M	30
Matched user-days (164M)	M	2.3M	70
Matched user-days (164M)	M	450K	365
Matched user-days (510M)	M	3.9M	130
Matched user-days (510M)	M	1.7M	300
Matched user-days (510M)	M	1.13M	450

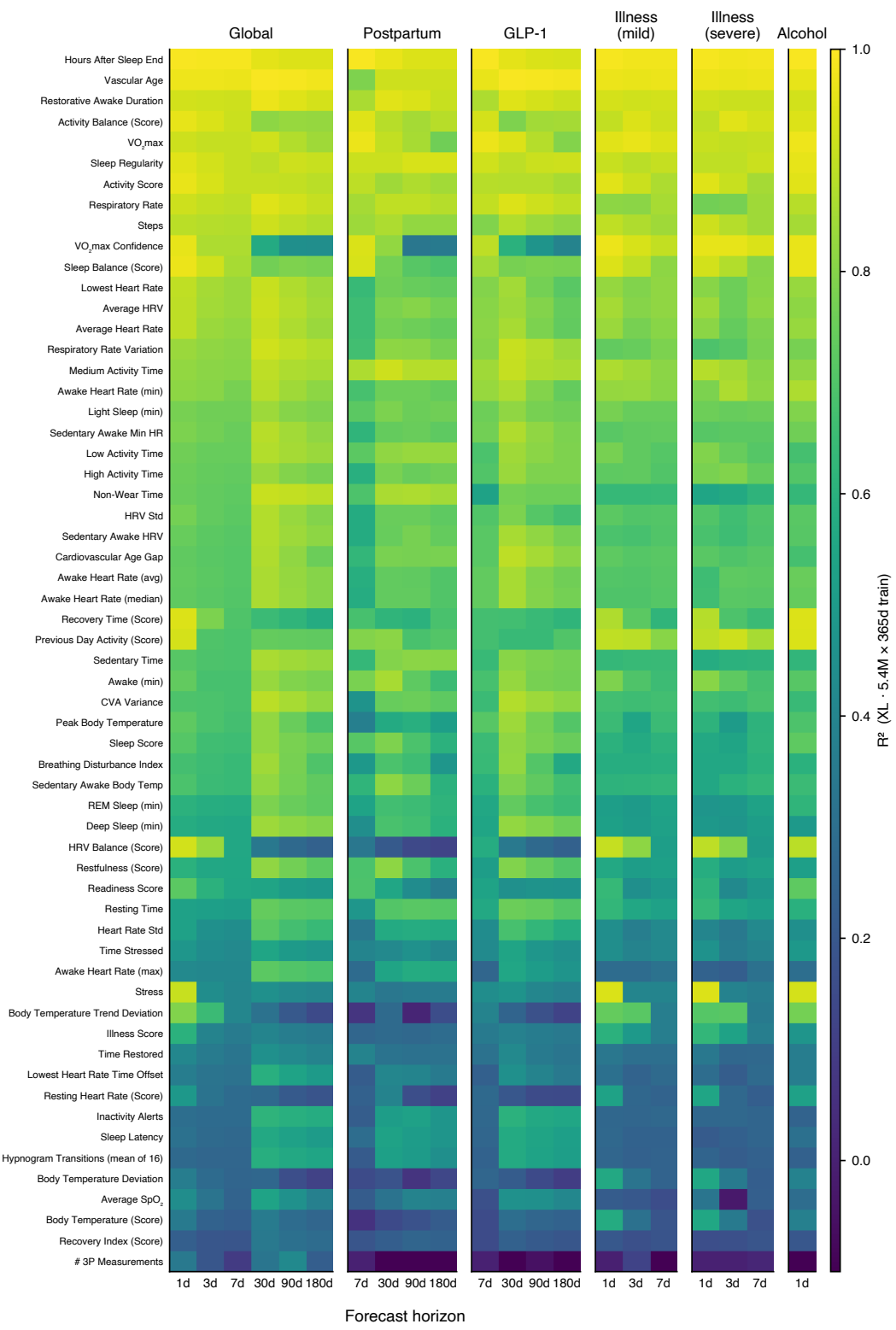

Figure 6: Results for XL model across all forecasting targets across all horizons, but globally and in specific cohorts. The base model is strong across most targets, with poor performance in highly sparse or difficult to predict outcomes such as SpO2 and derived scores around temperature. Hypnogram transitions (mean of 16) aggregates the performance on forecasting the 16 sleep-stage transition targets.

Table 7: Subject-level predictions for the flagship XS, M, and XL models (5.4M users, 365-day training context). All models show strong performance across all tasks, with the XL model performing best. Bolded results show significant improvements of the XL model over the smaller variants using a paired DeLong test.

Target	XS	M	XL	Pos. rate	n_{test}
Type-2 diabetes	0.883	0.888	**0.893**	0.024	135,696
Coronary heart disease	0.874	0.878	0.878	0.010	135,696
Blood-pressure medication	0.834	0.847	**0.852**	0.126	135,109
Hypertension	0.823	0.831	**0.836**	0.138	135,696
Sleep apnea	0.821	0.827	**0.833**	0.071	135,696
Stroke	0.779	0.784	0.787	0.007	135,696
High cholesterol	0.725	0.731	**0.733**	0.155	135,696
Atrial fibrillation	0.687	0.703	**0.713**	0.040	135,696
Cardiovascular family history	0.597	0.606	**0.608**	0.483	135,696

Table 8: Forecasting probes for the flagship XL model. Each row is a prediction at a future horizon of h days. AUROC is scored per-window; positive rate and window counts are reported at the window level, and the subject count reports how many distinct people contributed windows.

Target	AUROC	Pos. rate	n_{windows}	n_{subjects}
Forecasting Sickness Probes				
Sickness, h=1 day	.839	.359	31,599	7,431
Sickness, h=3 days	.827	.358	31,536	7,434
Sickness, h=7 days	.823	.355	31,368	7,433
Forecasting Cycle Phase Probes				
Cycle phase, h=1 day	.954	.405	8,945	3,347
Cycle phase, h=7 days	.931	.404	8,905	3,331

Table 9: Positioning dominates at both matched user-day budgets. Median ΔR^2 across the marquee grid. Q_1 (encoder quality) compares broad and deep cells at matched 30-day inference context; Q_2 (positioning) is the further gain when the deep cell reads its full context. Budgets: 164M user-days (5.4M$\times$30d vs. 450K$\times$365d) and 510M (3.9M$\times$130d vs. 1.13M$\times$450d). Values are medians of per-panel differences.

	164M user-days		510M user-days	
Cohort	Q_1	Q_2	Q_1	Q_2
Global	+.004	+.007	+.001	+.007
Alcohol	+.008	+.015	+.003	+.015
Illness (mild)	+.011	−.001	+.002	+.001
Illness (severe)	+.010	−.007	−.001	−.008
GLP-1	−.007	+.013	−.004	+.018
Postpartum	−.001	+.067	−.015	+.081

Table 10: Subject-level predictions for the flagship XL model (5.4M users, 365-day training context) when ridge probes are fit on validation users and scored on held-out test users. Though this probe setup results in limited positive examples, the results are in line with the main probe results and show the strength of our learned models on unseen labels.

Target	AUROC	Pos. rate	n_{test}
Subject-Level Probes			
Coronary heart disease	.858	.010	2,966
Type-2 diabetes	.835	.015	2,966
Sleep apnea	.816	.062	2,966
Blood-pressure medication	.791	.097	2,955
Hypertension	.789	.112	2,966
High cholesterol	.704	.148	2,966
Stroke	.647	.005	2,966
Atrial fibrillation	.607	.034	2,966
Cardiovascular family history	.560	.493	2,966

Table 11: Effect of each scaling axis on downstream R^2, per cohort, reported as median (max) across the marquee target grid. Gains concentrate on event-anchored cohorts: on medians, postpartum exceeds the global row by roughly $2.4\times$ on training context and $3.8\times$ on model size; the user-count axis moves the global row near zero, so its ratio is effectively unbounded (postpartum $+0.013$ vs global $+0.000$). We report both statistics because the max reflects the single most-affected target and horizon; reading it as the typical gain overstates the effect several-fold.

Cohort	Training context (7 → 365d)	Model size (XS → XL)	User count (0.5 → 5.4M)
Global	0.032 (0.117)	0.017 (0.121)	0.000 (0.025)
Alcohol	0.046 (0.108)	0.053 (0.145)	0.006 (0.069)
Illness (mild)	0.040 (0.104)	0.046 (0.108)	0.003 (0.021)
Illness (severe)	0.041 (0.102)	0.051 (0.134)	0.004 (0.032)
GLP-1	0.032 (0.095)	0.017 (0.058)	0.004 (0.020)
Postpartum	0.077 (0.145)	0.065 (0.214)	0.013 (0.056)

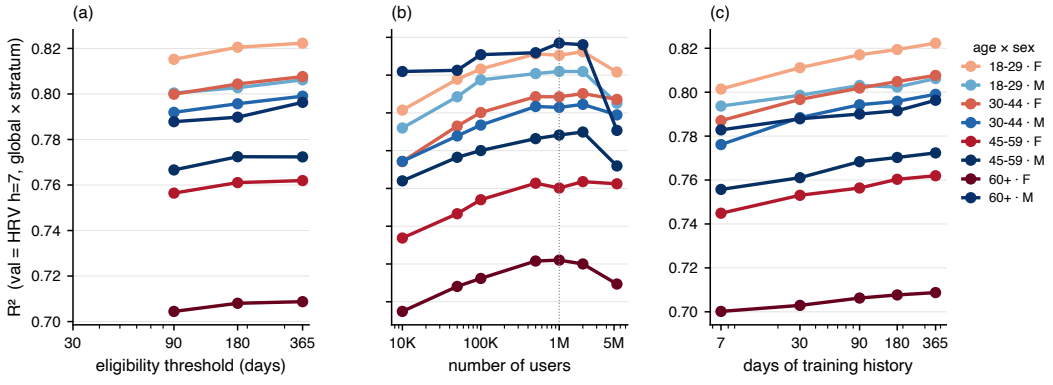

Figure 7: **The importance of depth reproduces across every age×sex subgroup.** Each panel replays Figure 2's data axis under the global cohort, with lines split by demographic subgroup, on `average_hrv` at h=7. (a) Trading users for per-user history lifts R^2 monotonically in all eight subgroups. (b) User count saturates near one million in every subgroup. (c) Per-user history keeps improving through 365 days in every subgroup. Older users start from a lower baseline (60+ female $R^2 \approx 0.71$ versus 18-29 female ≈ 0.83), while the shape of the depth curve is the same across subgroups.

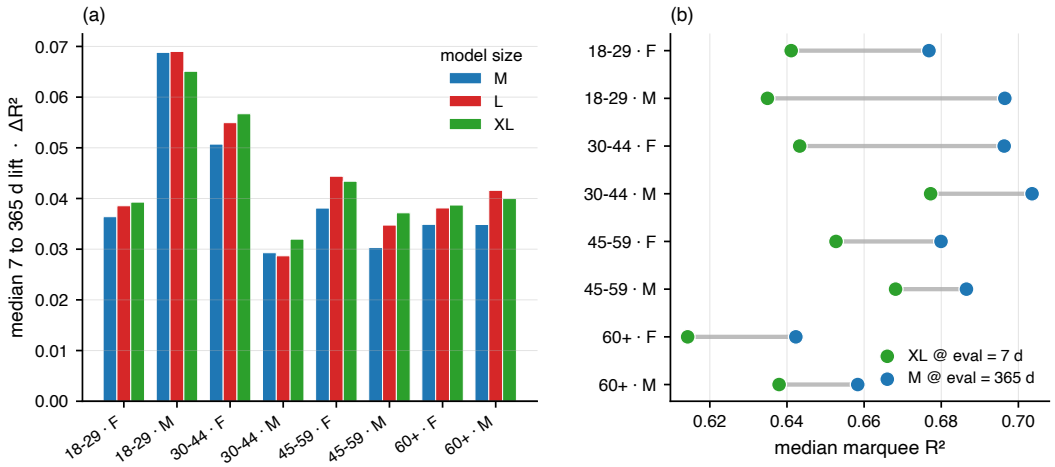

Figure 8: **Capacity multiplies context in every subgroup.** (a) Median 7 $\rightarrow$ 365-day training-context lift per subgroup $\times$ model size (M/L/XL) across the marquee grid. The L model's lift exceeds M's on 61% to 77% of marquee panels within every subgroup and XL's exceeds L's on 44% to 79%, with the endpoint ordering XL over M holding on 91% to 99%. (b) The XL model given only 7 days of inference context is beaten by the M model given a full year, on every one of the eight subgroups, with median-marquee gaps between -0.025 and -0.043 R^2. Older subgroups (60+) show smaller absolute gaps because their R^2 ceiling is lower, and the direction is preserved.

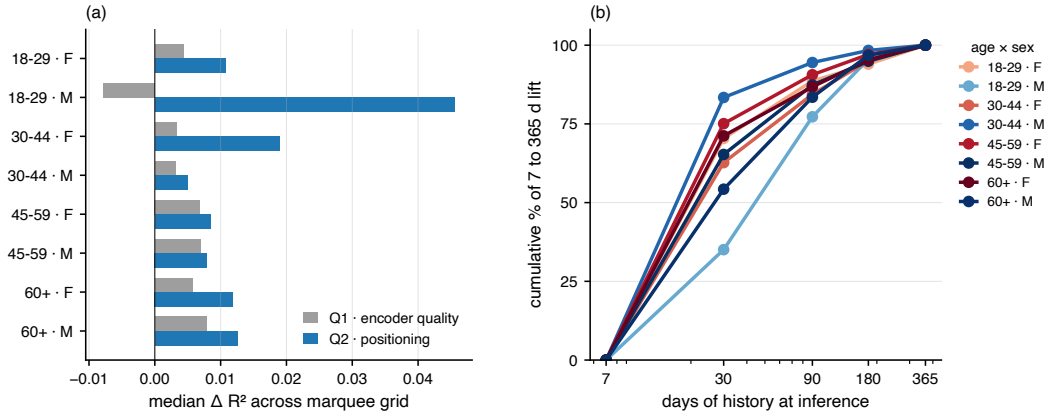

Figure 9: **Positioning dominates encoder quality in every subgroup.** (a) Q_1 (train-side, encoder-quality) versus Q_2 (eval-side, positioning) median ΔR^2 per subgroup, using the same decomposition as Figure 4. Q_2 exceeds Q_1 in every subgroup, by factors of 3.6$\times$ to 8.3$\times$. Two subgroups (18-29 female, 30-44 female) show the largest positioning-to-encoder gap matching the pattern the post-partum cohort showed, which is consistent with pregnancy/postpartum users concentrating in those bands. (b) Cumulative fraction of the 7 $\rightarrow$ 365-day eval-context lift as inference history grows, per subgroup. All subgroups saturate within 90 to 180 days; the fastest-saturating subgroup (18-29 female) captures 62% of the total lift by 30 days, while the slowest-starting subgroup (60+ female) reaches 55% by 30 days before climbing to 92% by 180 days.