Personalized coaching via coachable health age and counterfactual wearable trajectories

Simon A. Lee[1,2,*,†], Zeinab Esmaeilpour[1,*], Menglian Zhou[1], Marius Guerard[1], Sohrab Sidhu[1], Alexandru Solot[1], Hamid Palangi[1], Pramod Rudrapatna[1], Ming-Zher Poh[1], Max A. Xu[1], Daniel McDuff[1], Dimitrios Antos[1], Yuzhe Yang[1], Jeffrey N. Chiang[2], Anthony Z. Faranesh[1], Lindsey Sunden[1], Mark Malhotra[1], Shwetak Patel[1], A. Ali Heydari[1,‡], Ahmed A. Metwally[1,‡]

[1]Google Research, [2]University of California, Los Angeles, [*]Co-first, [†]Work Completed During Research Internship, [‡]Equal Leadership

arXiv:2611.08568v1 [q-bio.QM] 5 Nov 2026

Personalized health coaching requires moving beyond retrospective wearable summaries toward quantitative, actionable behavioral prescriptions. Here, we introduce Coachable Health Age (CHA), an interpretable health-age metric constructed from modifiable physiological and behavioral signals passively captured by consumer wearables. Across the WEAR-ME (N = 2,780), *All of Us* (N = 22,748), and longitudinal Fitbit (N = 1,464) cohorts, CHA aligns with established cardiometabolic health and risk scores and it improves prognostic discrimination for incident cardiovascular and metabolic disease beyond age, sex, and body mass index. To enable prospective behavioral optimization, we developed Sensor World Model (SensorWM), an action-conditioned wearable world model pretrained on over 2.2 million person-weeks of longitudinal sensor data from approximately 100,000 participants. SensorWM forecasts multi-week physiological trajectories conditioned on behavioral actions (R^2 = 0.664), outperforming a general-purpose time-series foundation model while preserving clinically relevant information for downstream prediction tasks, including diabetes and hypertension. Finally, we integrate CHA and SensorWM into an agentic counterfactual coaching system to synthesize personalized 12-week coaching plans. In blinded paired reviews, certified lifestyle coaches preferred counterfactual-grounded plans over plans grounded in wearable history alone for overall quality (71.4%, $P < 0.001$) and anticipated adherence (71.0%, $P < 0.001$). Board-certified clinicians endorsed safe exercise-intensity limits in 99.2% of ratings of counterfactual-grounded plans and 97.6% of ratings of comparison plans. Automated evaluations across the full cohort also favored counterfactual-grounded plans. These findings establish a framework integrating longitudinal wearable tracking, action-conditioned physiological forecasting, and personalized coaching, supporting prospective evaluation of its effects on behavioral adherence and cardiometabolic health.

1. Main

A substantial proportion of the global burden of chronic disease, particularly cardiometabolic disease, is attributable to modifiable behaviors such as physical inactivity, poor sleep, and unhealthy diet, and is therefore preventable [1, 2]. Translating this preventive potential into sustained behavior change requires measures of health that are personalized, intuitive, and responsive to changes in daily behavior. Although chronological age is among the strongest predictors of mortality and chronic disease, it fails to capture the substantial heterogeneity in physiological function among individuals of the same age. Biological or health age measures have therefore been developed from molecular, clinical and wearable-derived features to quantify deviations from expected age-related physiology [3, 4, 5]. By expressing multidimensional physiological information on a familiar age scale, these measures can make health status more interpretable and personally salient [6]. Related constructs such as “heart age” and “fitness age” have been used to communicate risk in accessible terms[7] and in some trials such framing has improved risk perception and health behaviors, including smoking cessation [8], although effects have been

Correspondence to: {aliheydari, aametwally}@google.com.
© 2026 Google. All rights reserved

inconsistent [7, 9]. Tracked longitudinally, health age could additionally serve as a feedback signal linking daily behaviors to longer-term changes in physiological state.

Most existing approaches derive health age by training machine-learning models to predict chronological age from multiple clinical biomarkers, or continuous wearable data [10, 11, 5], with a predicted age below chronological age, a negative age delta, interpreted as evidence of biological youthfulness or more favorable functional health. Second generation clocks trained on mortality or phenotypic outcomes improve risk predictions [4, 12], but like age trained models, they are optimized to capture risk rather than modifiability. Chronological age prediction and health coaching are not completely aligned objectives. A model optimized to predict chronological age or mortality risk may rely on physiological characteristics that are strongly age-associated but fixed, difficult to modify, or only weakly informative about behaviors an individual can act upon. They are also typically opaque; post hoc attribution methods can indicate which inputs influenced an estimate, but not how that estimate would change if a person altered their behavior[13]. A health-age measure intended for coaching should therefore be built from interpretable, repeatedly measurable characteristics that can serve as practical targets for behavioral feedback [14, 15, 16].

A coachable health-age measure describes an individual's current state. Prospective coaching additionally requires understanding how that state may evolve by comparing health trajectories under alternative behavioral scenarios [17]. The density and continuity of consumer-wearable signals make them well suited to this task. Recurrent and transformer models have been used to model physiological states and incident risk from longitudinal health records and wearable data [18, 5]. This work has culminated in large-scale wearable "foundation models" trained using self-supervised objectives across thousands to millions of participants, which learn general-purpose physiological representations [19, 20, 21, 22, 23, 24, 25] and support imputation and interpolation of sensor values over time [26]. In parallel, the digital-twin paradigm proposes individualized virtual replicas that simulate personal health trajectories to enable anticipatory care [27, 28]. These approaches are predominantly observational. They estimate the future associated with observed or persisting behavioral patterns. Related longitudinal outcome-modeling methods have considered predictions under alternative treatment sequences [29, 30, 31, 32], while action-conditioned world models provide a general computational framework for representing how predicted future states depend on specified actions [33, 34]. For wearable health coaching, however, the relevant actions are not primarily discrete clinical treatments, but continuously varying behaviors such as physical activity, sleep, and cardiovascular exercise. This motivates an action-conditioned wearable sensor model that can generate individualized physiological projections under feasible behavioral scenarios and summarize those projections using a common, coachable outcome.

Translating such projections into guidance that people can follow is a further challenge. Large language models (LLMs) can deliver personalized, conversational health coaching and have shown expert-level performance in interpreting wearable sleep and fitness data [35, 36, 37]. However, current LLM-based health coaches are typically grounded in retrospective summaries of an individual's data rather than in quantitative projections of how their physiology would respond to specific behavioral changes, yielding recommendations that rely on qualitative population heuristics rather than calibrated, individualized dose-response estimates[36, 35]. Moreover, LLMs can generate plausible but unsupported claims [38], and despite recent benchmarks in wearable-based coaching [36, 37], few systems have been rigorously evaluated by both lifestyle-coaching and clinical experts [39]. Coupling LLM coaching with an explicit model of physiological response and an interpretable outcome measure could make recommendations both personalized and directly traceable to their projected health benefits.

Here, we introduce **Coachable Health Age (CHA)**, an interpretable health-age measure built from physiological and behavioral signals that are passively and repeatedly measured by consumer

wearables, including resting heart rate, sleep duration and physical activity, and that reflect modifiable aspects of everyday health. We embed CHA within an architecture that couples an action-conditioned wearable **Sensor World Model (SensorWM)** with an agentic coaching system. The world model learns predictive physiological dynamics from longitudinal wearable trajectories and generates model-implied future states conditioned on alternative patterns of daily activity, sleep, and structured exercise. The coaching agent integrates these counterfactual projections with a feasibility-constrained optimization procedure to produce personalized behavioral plans grounded in projected changes in CHA. We evaluate each component of the framework, assessing the clinical, prognostic, and longitudinal validity of CHA, the forecasting fidelity of SensorWM in driving the coaching simulation engine, the behavior of its scenario-conditioned projections, and the quality of coaching plans generated with and without simulation-derived information. In blinded evaluations, lifestyle coaches consistently preferred plans informed by CHA and SensorWM projections over plans grounded in baseline wearable data alone for overall plan quality and anticipated long-term adherence, and clinicians rated both sets of plans similarly on exercise-intensity safety and physiological accuracy.

2. Results

2.1. Study design and dataset characteristics

We designed the study as three linked analyses evaluating the core components of our end-to-end CHA and counterfactual coaching framework (Figure 1; cohort demographics and baseline wearable characteristics in Table 1). First, we constructed and evaluated CHA, an age-equivalent summary of wearable-derived physiological and behavioral health built from repeatedly measurable, modifiable signals (Figure 1a). We assessed its alignment against established cardiovascular and metabolic health and risk scores (modified Life's Essential 8 and Metabolic Syndrome Severity) and self-reported health in the WEAR-ME cohort, and characterized its within-person longitudinal modifiability across repeated 3-month windows over two years in the longitudinal Fitbit cohort. Independent prospective clinical validation was then performed in the *All of Us* Research Program by relating baseline and longitudinal ΔCHA (ΔCHA = CHA – chronological age) to incident cardiometabolic disease, specifically hypertension, type 2 diabetes, and cardiovascular disease as well as incremental prognostic discrimination beyond age, sex, and body mass index (Figure 1b).

Second, we developed SensorWM, an action-conditioned latent world model for longitudinal wearable data (Figure 1c). SensorWM was pretrained on real world longitudinal wearable trajectories from 96,957 participants (approximately 100,000), comprising 2.28 million total person-weeks of wearable data (182,797 valid 12-week windows; 42,039 person-years). The model was trained to represent physiological state in a latent space and to predict its future evolution conditional on both recent physiological trajectory and observed behavioral actions. Model fidelity was evaluated by comparing generated and observed physiological trajectories, feature-level forecasting accuracy, and preservation of clinically relevant information on downstream health prediction tasks in a held-out WEAR-ME cohort.

Third, we evaluated whether grounding an LLM coaching agent in CHA and SensorWM counterfactual projections improved personalized 12-week behavioral prescriptions (Figure 1d,e). For each eligible WEAR-ME participant, SensorWM was initialized from recent wearable history to simulate future physiological trajectories under candidate modifications in daily steps, sleep duration, and cardiorespiratory exercise. Candidate trajectories were mapped to projected changes in CHA and ranked using a feasibility-constrained optimization objective that balances projected CHA reduction against participant-specific behavioral effort and safety bounds (Figure 1d). We then compared two matched LLM coaching conditions that shared the same underlying language model, participant wearable history, and plan structure, differing only in access to simulation-

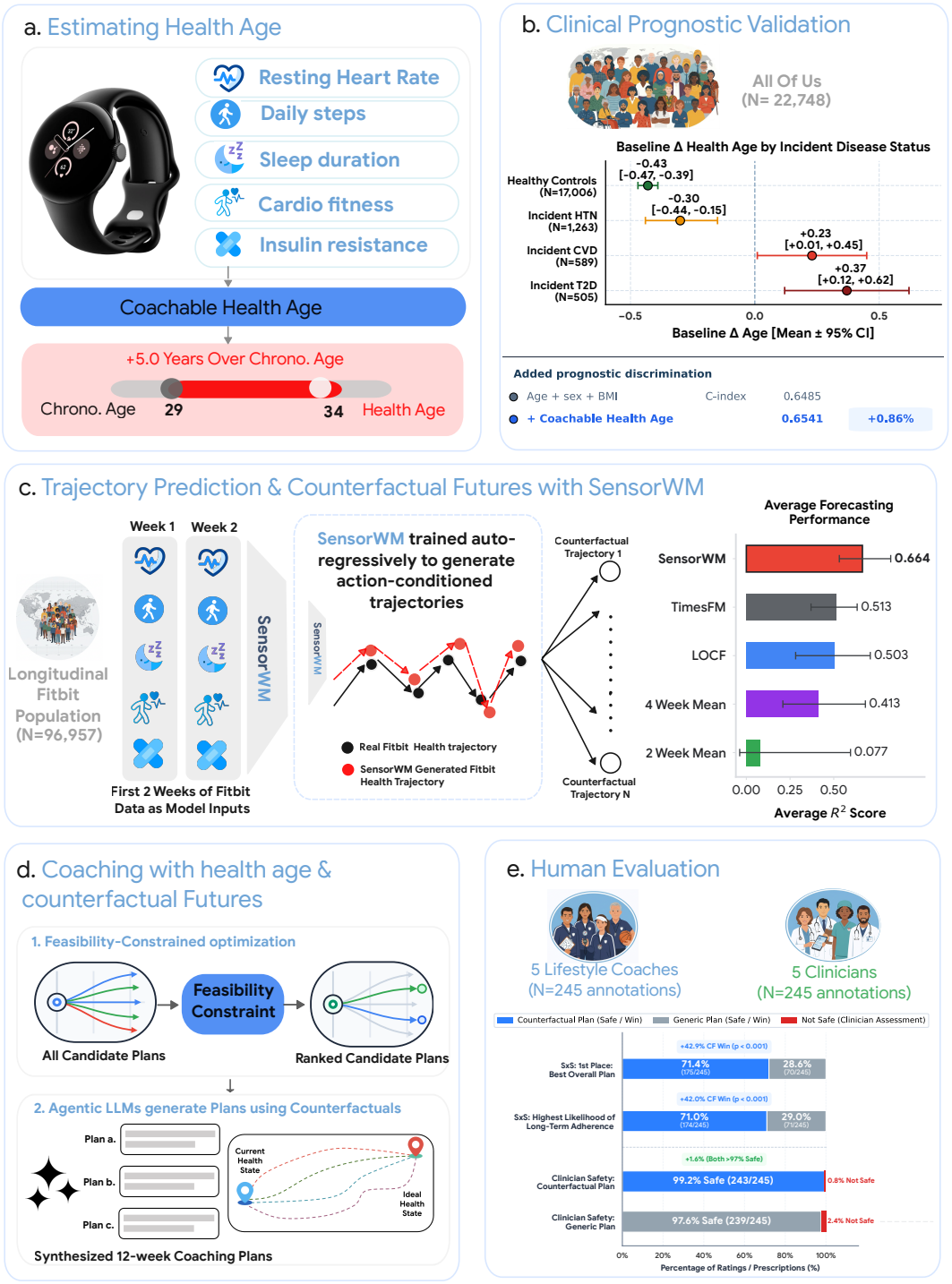


Figure 1 | **Coachable Health Age (CHA) framework for personalized health optimization. (a)** Wearable sensor physiological and behavioral metrics are integrated to estimate CHA relative to chronological age. **(b)** Clinical prognosis across incident cardiometabolic outcomes in All of Us. **(c)** SensorWM auto-regressively simulates action-conditioned counterfactual lifestyle trajectories over a 12-week horizon. **(d)** Feasibility-constrained optimization balances forecasted Health Age reduction against historical effort penalties and clinical safety gates to synthesize a personalized 12-week coaching plan. **(e)** Dual-track human evaluation by lifestyle coaches (10 behavioral rubrics and side-by-side preference) and clinicians (clinical safety gate and 6 clinical rubrics).

derived information: (1) **Counterfactual Coaching**, which received the participant's baseline CHA, projected status-quo trajectory, and feasibility-ranked counterfactual pathways, and (2) **Generic Coaching**, which generated recommendations from observed baseline wearable history alone. Plan quality, clinical safety, physiological plausibility, and anticipated long-term adherence were evaluated through a blinded, randomized dual-track human expert study with five certified lifestyle coaches and five board-certified clinicians across stratified WEAR-ME personas, complemented by a full-cohort automated evaluation across the WEAR-ME dataset using rubric-aligned LLM autoraters (Figure 1e). All plan generation and expert evaluations were conducted offline using retrospective participant profiles prior to prospective deployment.

2.2. Coachable Health Age Framework

To serve as both a continuous summary of everyday health and a quantitative target for prospective behavioral coaching, a health-age measure must satisfy seven core design criteria (Supplementary Table 2): it must be (1) *predictable from consumer wearables* via passive, high-frequency sensing; (2) *explainable* at the level of individual inputs; (3) *flexible and extensible* to additional features; (4) *engageable* for longitudinal user communication; (5) *coachable*, mapping directly to modifiable daily habits; (6) *movable* within individuals over coaching relevant timescales; and (7) *grounded* in clinical outcomes. Existing age formulations satisfy only subsets of these requirements. Chronological age provides an intuitive reference scale but cannot capture physiological heterogeneity among peers of the same age. Machine learning clocks trained to predict chronological age from wearable streams (such as foundation-model embeddings) capture age correlated sensor variation, but their training objective does not prioritize modifiable behaviors or present cardiometabolic health, and their black-box representations limit actionable attribution. Conversely, clinical biological clocks such as PhenoAge are grounded in mortality-associated blood biomarkers, but rely on invasive, episodic laboratory draws that cannot be tracked continuously by consumer wearables. Finally, directly prompting large language models (LLMs) to estimate a numerical health age from wearable summaries introduces stochastic variation across inference settings and lacks deterministic epidemiological calibration (evaluated separately in Supplementary Information A.1: LLM Health Age & Consistency).

We therefore formulated Coachable Health Age (CHA) by mapping repeatedly measurable, modifiable wearable signals, core behaviors and physiology comprising resting heart rate, daily steps, sleep duration, and weekly Active Zone Minutes (AZM)—onto an intuitive age-equivalent scale using collinearity-adjusted epidemiological hazard ratios (Figure 2a; Methods). Two architectural choices distinguish this framework. First, rather than parameterizing risk solely against late-life all-cause mortality, CHA anchors its core wearable domains to published epidemiological associations with cardiometabolic outcomes, including type 2 diabetes, heart failure, and cardiovascular disease, emphasizing risk factors that are measurable through consumer wearables and potentially responsive to behavioral modification. Second, because CHA aggregates risk additively in log-hazard space rather than through a monolithic neural network, the framework is inherently modular and expandable: optional wearable or clinical signals such as cardiorespiratory fitness (VO_2 max), sleep midpoint regularity, and wearable-estimated insulin resistance, or emerging sensing modalities can be seamlessly incorporated when available without retraining the underlying formulation. Together, these design characteristics motivate CHA as an interpretable, cardiometabolic grounded outcome metric (Supplementary Table 2) and reflect design considerations rather than an empirical ranking of alternative age measures.

2.3. Cardiometabolic prognosis and longitudinal modifiability of CHA

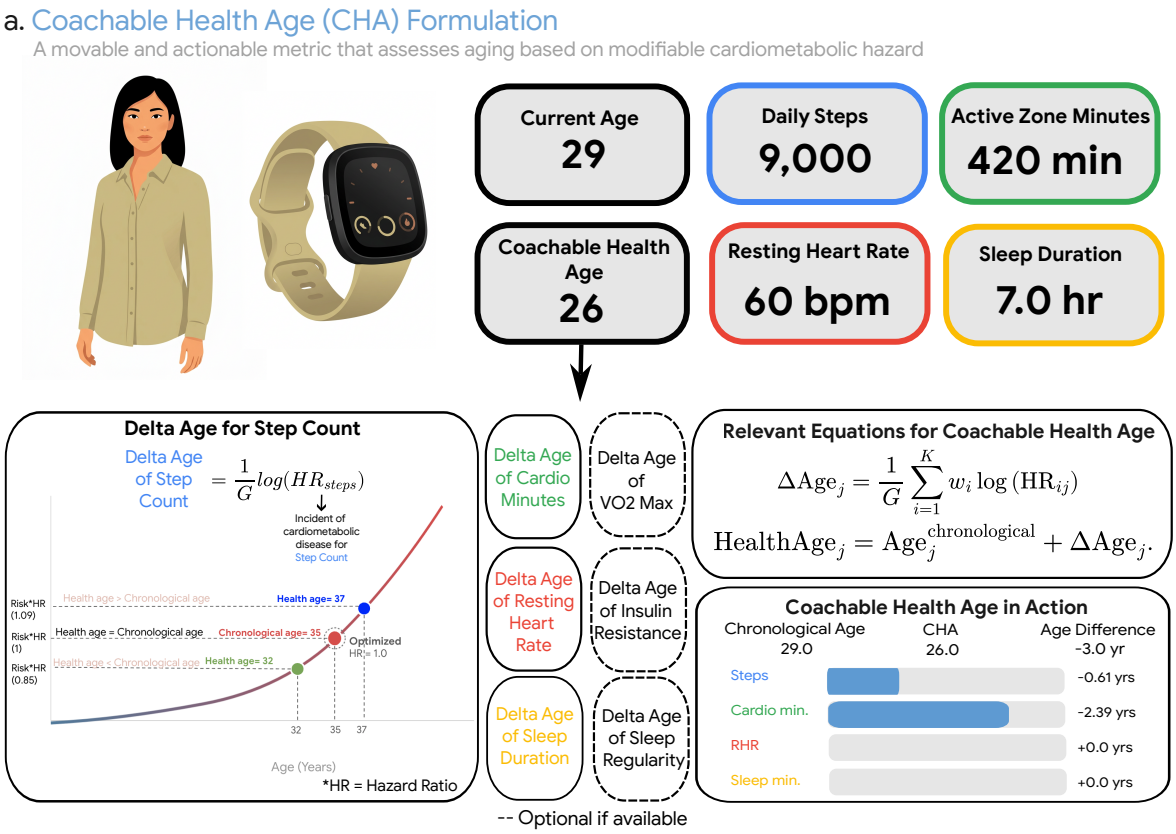
Because CHA incorporates modifiable lifestyle and physiological domains emphasized in preventive cardiometabolic disease, we first verified its alignment against established clinical health

Table 1 | **Demographic, longitudinal wearable, and clinical characteristics of the study cohorts.** Categorical variables are reported as *n* (%) and continuous variables as mean ± standard deviation, unless otherwise specified. The disease-free baseline cohort excludes participants with a history of hypertension, type 2 diabetes, or cardiovascular disease on or before the end of their 90-day baseline wearable window.

	CHA AoU Analysis			SensorWM / Expert Evaluation	
Characteristic	**Total Wearable Cohort**	**Disease-Free Baseline (KM)**	**Cox / C-Index Cohort**	**Pre-training Cohort**	**WEAR-ME Evaluation**
Cohort Size and Wearable Observation					
Participants, *N*	22,748	18,423	14,663	96,957	2,780
Total valid 3-month wearable windows, *n*	276,378	227,557	188,033	182,797	3,289
Valid 3-month windows per participant	12.1 ± 8.3	12.4 ± 8.4	12.8 ± 8.5	1.9 ± 1.5	1.2 ± 0.4
Active wearable tracking, person-years	68,101	56,071	46,332	42,039	756
Follow-up span, years	3.59 ± 2.28	3.65 ± 2.30	3.80 ± 2.31	0.43 ± 0.35	0.38 ± 0.08
Baseline Demographics					
Age, years	50.2 ± 15.5	48.2 ± 15.4	48.7 ± 15.5	40.7 ± 16.6	48.5 ± 12.0
Age range, years	13–85	13–84	13–84	18–100	23–100
Sex, *n* (%)					
Female	16,031 (70.5)	13,164 (71.5)	10,446 (71.2)	47,651 (49.2)	1,652 (59.4)
Male	6,209 (27.3)	4,814 (26.1)	3,872 (26.4)	49,294 (50.8)	1,123 (40.4)
Other / Unknown / Not reported	508 (2.2)	445 (2.4)	345 (2.4)	12 (0.0)	5 (0.2)
Body Mass Index (BMI)					
BMI, kg/m^2	29.61 ± 7.18[a]	28.95 ± 6.88[b]	28.95 ± 6.88	26.21 ± 6.46	29.52 ± 6.85
Baseline Wearable and Physiological Characteristics					
Daily steps, steps/day	7,846 ± 3,574	8,040 ± 3,560	8,131 ± 3,511	5, 604 ± 3, 601	7, 405 ± 3, 519
Resting heart rate proxy, bpm	55.2 ± 7.3	55.1 ± 7.2	55.2 ± 7.1	69.1 ± 8.8	68.2 ± 8.6
Main sleep duration, hours/night	6.36 ± 1.20	6.41 ± 1.17	6.43 ± 1.16	6.75 ± 1.01	7.12 ± 1.15
Active Zone Minutes, minutes/week	373.7 ± 374.5	381.3 ± 375.7	388.1 ± 377.2	302.4 ± 303.8	198.4 ± 142.6
Coachable Health Age					
Coachable Health Age, years	49.97 ± 15.84	47.77 ± 15.64	48.18 ± 15.69	44.10 ± 17.13	48.46 ± 12.02
Health Age Delta, Δ years	−0.26 ± 2.67	−0.42 ± 2.60	−0.51 ± 2.56	+3.40 ± 3.80	−0.04 ± 3.00
Cardiometabolic Disease Status					
Pre-existing cardiometabolic condition, *n* (%)	4,325 (19.0)	0 (0.0)	0 (0.0)	—	—
Incident composite diagnosis, *n* (%)[c]	1,417 (6.2)	1,417 (7.7)	1,402 (9.5)	—	—
Incident hypertension	1,263 (5.6)	1,138 (6.2)	1,125 (7.7)	—	—
Incident type 2 diabetes	505 (2.2)	315 (1.7)	309 (2.1)	—	—
Incident cardiovascular disease	589 (2.6)	309 (1.7)	304 (2.1)	—	—

Notes. Values are *n* (%) or mean ± standard deviation unless otherwise specified. Percentages are calculated using the corresponding cohort size and may not sum to 100% because of rounding. An em dash (—) indicates that a characteristic was not available for the corresponding cohort. Observation duration in the pre-training and WEAR-ME evaluation cohorts is measured in weeks, whereas longitudinal clinical follow-up is summarized using valid 3-month wearable windows and years. In the *All of Us* cohort, all participants were ≥ 18 years of age at study enrollment; baseline age reflects age at the first valid 90-day wearable window, which includes retrospective Bring Your Own Device (BYOD) Fitbit records collected from age 13 onward. The Cox / C-index cohort is restricted to disease-free participants with available EHR-derived BMI. [a] BMI available for N = 18,757 participants in the total wearable cohort. [b] BMI available for N = 14,663 participants in the disease-free cohort. [c] Individual incident disease counts are not mutually exclusive, as participants may develop more than one condition during follow-up; in the total wearable cohort, condition-specific incident counts exclude only baseline history of that respective condition, whereas in the disease-free and Cox cohorts all participants are free of all three conditions at baseline. In the pre-training cohort, 2.28 million total person-weeks of observations yielded 182,797 complete 12-week (84-day) windows (2.19 million windowed person-weeks; 42,039 person-years).

and risk scores namely modified Life's Essential 8 (mod-LE8) and Metabolic Syndrome Severity Score (MSSS) as well as independent self-reported health status in the WEAR-ME cohort (Supplementary Figure 1). Whereas wearable models trained to predict chronological age showed weak concordance with current cardiometabolic health (mod-LE8 $|r| = 0.141$; MSSS $|r| \leq 0.048$) and direct LLM-based health-age estimates were poorly calibrated ($|r| = 0.110$–0.214), CHA preserved strong alignment with both mod-LE8 ($|r| = 0.512$) and MSSS ($|r| = 0.355$), exceeding blood-based PhenoAge ($|r| = 0.406$ and $|r| = 0.319$, respectively) and varying monotonically across self-reported health categories from excellent to poor. Thus, unlike wearable clocks, CHA



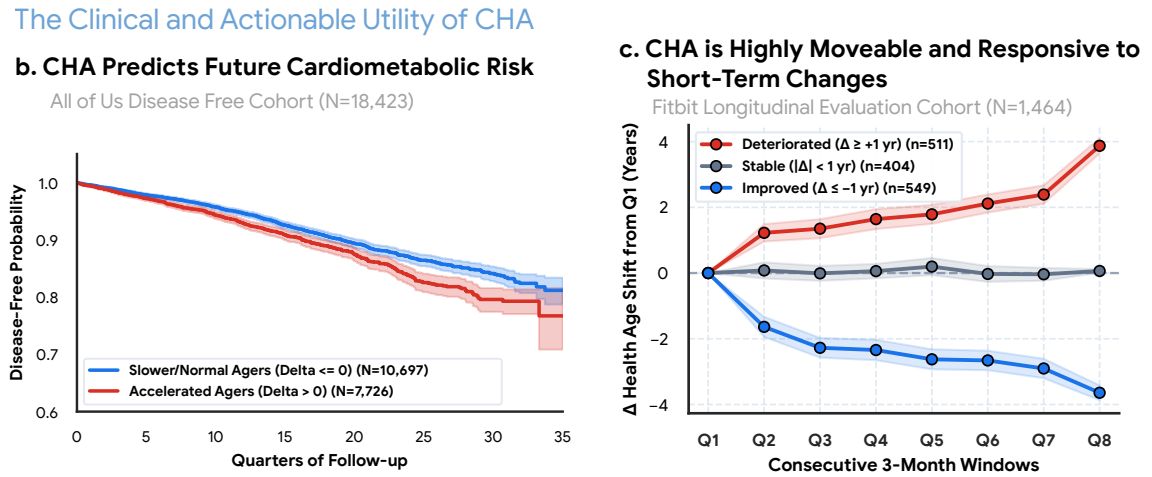


Figure 2 | **Formulation, cardiometabolic prognosis, and short-term modifiability of CHA. (a)** Overview of the CHA formulation, mapping core wearable signals (resting heart rate, sleep duration, cardiorespiratory minutes, and step count) and optional metabolic or fitness markers (insulin resistance, sleep regularity, and VO_2 max) onto a Gompertz cardiometabolic hazard curve (*left*), alongside the governing equations and an illustrative four-component calculation (*right*). The hypothetical example assumes a chronological age of 29 years, 9,000 daily steps, 420 weekly Active Zone Minutes, a resting heart rate of 60 beats per minute, and 7 hours of nightly sleep. Optional components are omitted. **(b)** Kaplan–Meier estimates of remaining free of incident cardiometabolic disease in the *All of Us* cohort (N = 18,423), stratified by baseline Health Age Delta (ΔCHA $\leq$ 0, n = 10,697 versus ΔCHA > 0, n = 7,726). **(c)** Mean ΔCHA shift from baseline (Q1) across eight consecutive 3-month windows (2 years) among deteriorated ($\Delta \geq +1$ yr, n = 511), stable ($|\Delta| < 1$ yr, n = 404), and improved ($\Delta \leq -1$ yr, n = 549) trajectory groups (N = 1,464). Shaded bands indicate 95% confidence intervals.

directly reflects an individual's current cardiometabolic and perceived health.

To independently validate the prognostic utility of CHA, we evaluated longitudinal incident cardiometabolic disease in the *All of Us* Research Program wearable cohort (N = 22,748, of whom N = 18,423 were free of all three conditions at baseline and N = 4,325 had pre-existing baseline disease; Figure 1b). Compared with healthy controls who remained disease-free throughout follow-up (N = 17,006; mean ΔCHA = −0.43 years, 95% CI : $[-0.47, -0.39]$), baseline ΔCHA was higher among participants who subsequently developed incident hypertension (N = 1,263; −0.30 years, $[-0.44, -0.15]$), cardiovascular disease (N = 589; +0.23 years, $[+0.01, +0.45]$), or type 2 diabetes (N = 505; +0.37 years, $[+0.12, +0.62]$; Figure 1b). Over up to 36 quarters of follow-up, Kaplan–Meier analyses confirmed that baseline accelerated agers (ΔCHA $>$ 0, n = 7,726) experienced significantly lower cardiometabolic disease-free survival than slower or normal agers (ΔCHA $\leq$ 0, n = 10,697; 8-year survival 79.28% versus 82.41%, log-rank $P = 7.09 \times 10^{-6}$; Figure 2b and Extended Data Table 1). In multivariable Cox proportional hazards models adjusted for baseline chronological age, sex, and BMI, each 1-year increase in ΔCHA remained independently associated with higher incident cardiometabolic risk both at baseline (HR = 1.02, 95% CI : $[1.01, 1.04]$, $P = 2.8 \times 10^{-4}$; N = 14,663, 1,402 events) and in lagged time-varying models where ΔCHA_{t-1} predicted incident diagnosis in the subsequent quarter (HR = 1.04, 95% CI : $[1.02, 1.05]$, $P = 6.0 \times 10^{-7}$; N = 13,842, 164,336 person-quarters; Extended Data Table 2). Crucially, adding CHA to the clinical baseline of age, sex, and BMI significantly improved prognostic discrimination across the full cohort (Harrell's C-index 0.6541 versus 0.6485, ΔC = +0.0056, P = 0.005), in 5-fold cross-validation (ΔC = +0.0053, P = 0.016), and most pronouncedly in adults under 50 years of age (N = 7,399; C-index 0.6762 versus 0.6648, ΔC = +0.0114, $P < 0.001$; Extended Data Table 3 and Supplementary Figure 2).

2.4. Sensor world models learn physiological health trajectories

Estimating an actionable Health Age requires more than measuring an individual's present physiological state. It requires a model of how that state evolves over time and how predicted trajectories vary with observed behavioral patterns. We therefore developed SensorWM, an action-conditioned sensor world model trained to reconstruct and forecast longitudinal wearable measurements (Figure 3a). SensorWM was pretrained on trajectories from 96,957 research-consented participants, comprising 2.28 million total person-weeks. Each participant was represented as a weekly sequence of physiological states and observable actions. State inputs included cardiovascular (e.g., Heart Rate Variability), cardiopulmonary (e.g., Respiratory Rate), sleep (e.g., Deep sleep minutes), and variability-related (e.g. dispersion) measurements (e.g., Relative Sample SD30), whereas action inputs captured modifiable behaviors, including step count, sleep duration and time spent in cardio-intensity activity (Full Decomposition Breakdowns of Feature Splits in Supplementary Table 6).

SensorWM consists of a physiological state encoder and an action-conditioned transition model. A multilayer perceptron first encodes physiological measurements into a 128-dimensional latent representation. The transition model then predicts how this representation changes over time based on the current physiological state, recent changes in that state, and observed behavioral actions. Specifically, it learns a linear transition operator that updates the latent representation using its recent trajectory. Both components are trained jointly using a self-supervised joint-embedding predictive architecture (JEPA), closely following the training objective of LeWorldModel which models trajectory fidelity and a regularization term [40].

This state–action formulation enabled SensorWM to predict future physiological states based on both observed behaviors and recent physiological changes, allowing the model to capture not only an individual's current health status but also the direction and trajectory of their physiological evolution. The resulting architecture supports autoregressive forecasting and the simulation of alternative behavioral scenarios over extended horizons, providing the physiological projections

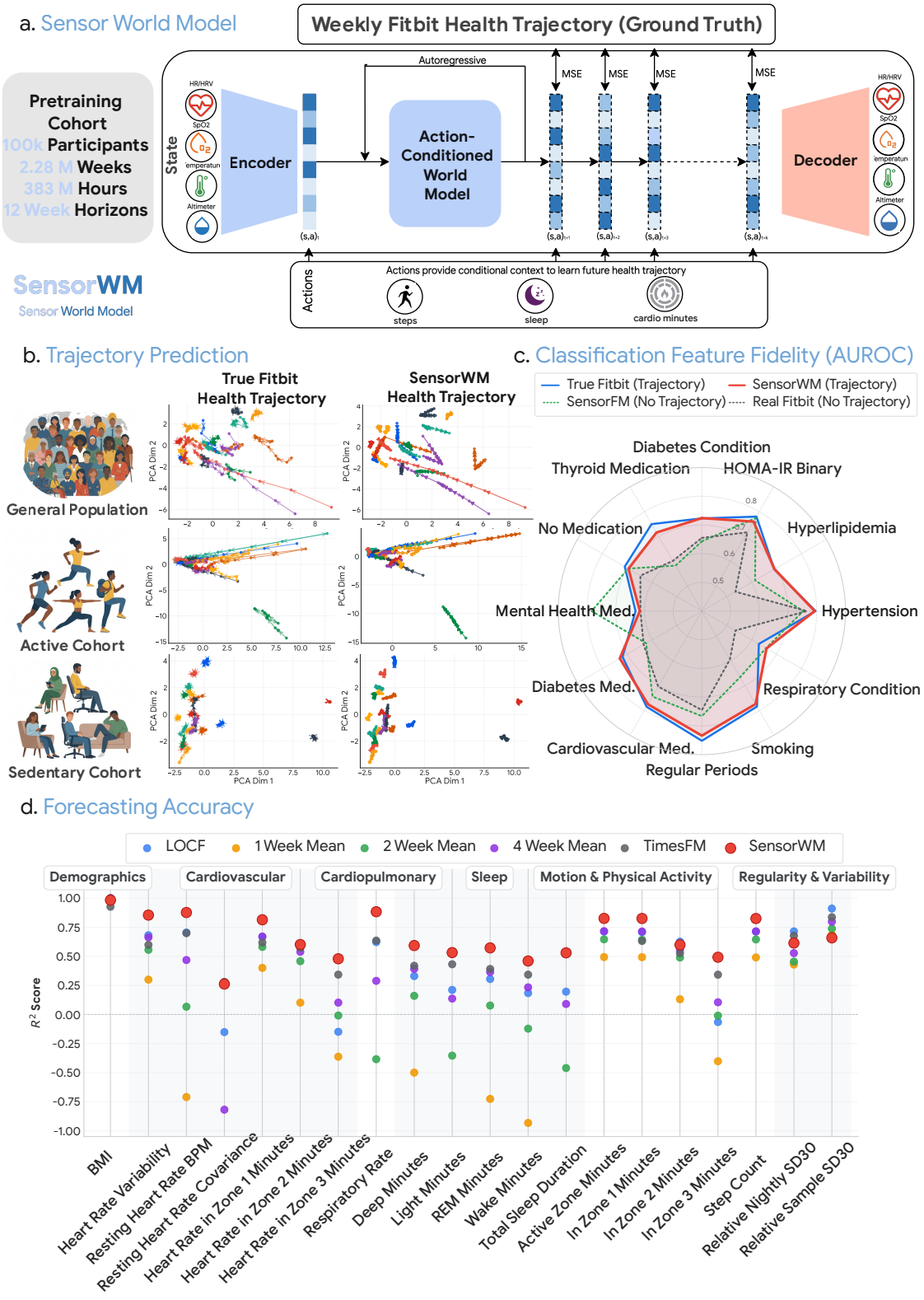


Figure 3 | **Workflow and Validation of the Sensor World Model. (a)** Sensor world model architecture for action-conditioned prediction of future health trajectories from wearable sensor data. **(b)** PCA projections of ground-truth (left) and SensorWM-predicted (right) health trajectories across general, active, and sedentary cohorts. **(c)** Downstream classification performance (AUROC) comparing real and SensorWM-generated features, with and without trajectory information. **(d)** Forecasting accuracy (R^2) across 20 wearable health features, comparing SensorWM against baseline models.

used in downstream counterfactual coaching. Additional details on model architecture, training objectives, and recursive inference are provided in Methods.

We next tested whether generated trajectories retained clinically relevant information. This evaluation is more stringent than minimizing point-wise reconstruction error; a forecast can be numerically close to the observed sequence while erasing weak but medically informative covariation across its learned latent representation. We therefore trained matched downstream classifiers using two settings: (1) observed wearable trajectories, (2) SensorWM-generated trajectories. We then evaluated the two settings across health conditions, laboratory-derived metabolic phenotypes, medication use, behavioral activity and women's health prediction tasks. Across most downstream prediction tasks, SensorWM-generated trajectories outperformed the pretrained Sensor Foundation Model (SensorFM) and Fitbit features without longitudinal trajectory information (Figure 3c). For example, generated trajectories yielded an AUROC of 0.724 for diabetes compared with 0.643 using SensorFM ($p = 1.85 \times 10^{-5}$), 0.692 versus 0.614 for hyperlipidemia ($p = 9.64 \times 10^{-10}$), and 0.793 versus 0.753 for hypertension ($p = 5.85 \times 10^{-4}$). We next evaluated whether SensorWM preserved the clinical information contained in the true longitudinal Fitbit trajectories. Across these tasks, classification performance from generated trajectories closely followed that obtained from the observed data (Figure 3c), with most comparisons showing no statistically significant differences. For example, generated trajectories yielded an AUROC of 0.724 for diabetes compared with 0.723 using Fitbit observed trajectories (p=0.7037), 0.692 versus 0.691 for hyperlipidemia, and 0.793 versus 0.796 for hypertension (p=0.7784). Performance was likewise preserved for HOMA-IR-defined insulin resistance, with an AUROC of 0.760 from generated trajectories compared with 0.780 from observed data (p=0.2782). These results suggest that SensorWM captures clinically informative longitudinal dynamics while preserving much of the predictive information present in observed wearable trajectories.

We then quantified forecasting accuracy across wearable domains and compared SensorWM with persistence and temporal baselines, including the mean of the preceding one, two, or four weeks, last observation carried forward (LOCF), and the time-series foundation model TimesFM [41]. SensorWM was initialized using the first two observed physiological states and subsequently propagated autoregressively, with each physiological state transition conditioned on the corresponding behavioral action vector. The largest improvements occurred for measurements with substantial temporal variation and cross-modal dependence. SensorWM attained $R^2 = 0.877$ for resting heart rate, compared with 0.705 for last observation carried forward and 0.686 for TimesFM. For heart-rate variability, SensorWM reached $R^2 = 0.855$, compared with 0.682 and 0.605, respectively. Respiratory rate was forecast with $R^2 = 0.884$, substantially exceeding last observation carried forward ($R^2 = 0.623$) and TimesFM ($R^2 = 0.503$). SensorWM also performed strongly for time spent in lower-intensity heart-rate zones, achieving $R^2 = 0.814$ for heart-rate zone 1 minutes and $R^2 = 0.826$ for activity zone 1 minutes. These features reflect both behavior and physiological response; predicting them therefore requires the model to integrate activity patterns with individual cardiovascular dynamics rather than rely solely on univariate persistence.

2.5. SensorWM enables individualized counterfactual simulation and feasibility-constrained coaching

We next used SensorWM to project how individual physiological trajectories may change under alternative behavioral scenarios. For each participant, SensorWM was initialized from the same recent physiological history and rolled forward for 12 weeks under candidate changes in daily steps, sleep duration, and cardio-intensity activity. Each projected trajectory was decoded into wearable-derived physiological measures and mapped to CHA, allowing alternative model-implied futures to be compared from a common starting state (Figure 4a).

To translate these projections into actionable coaching options, we coupled SensorWM with

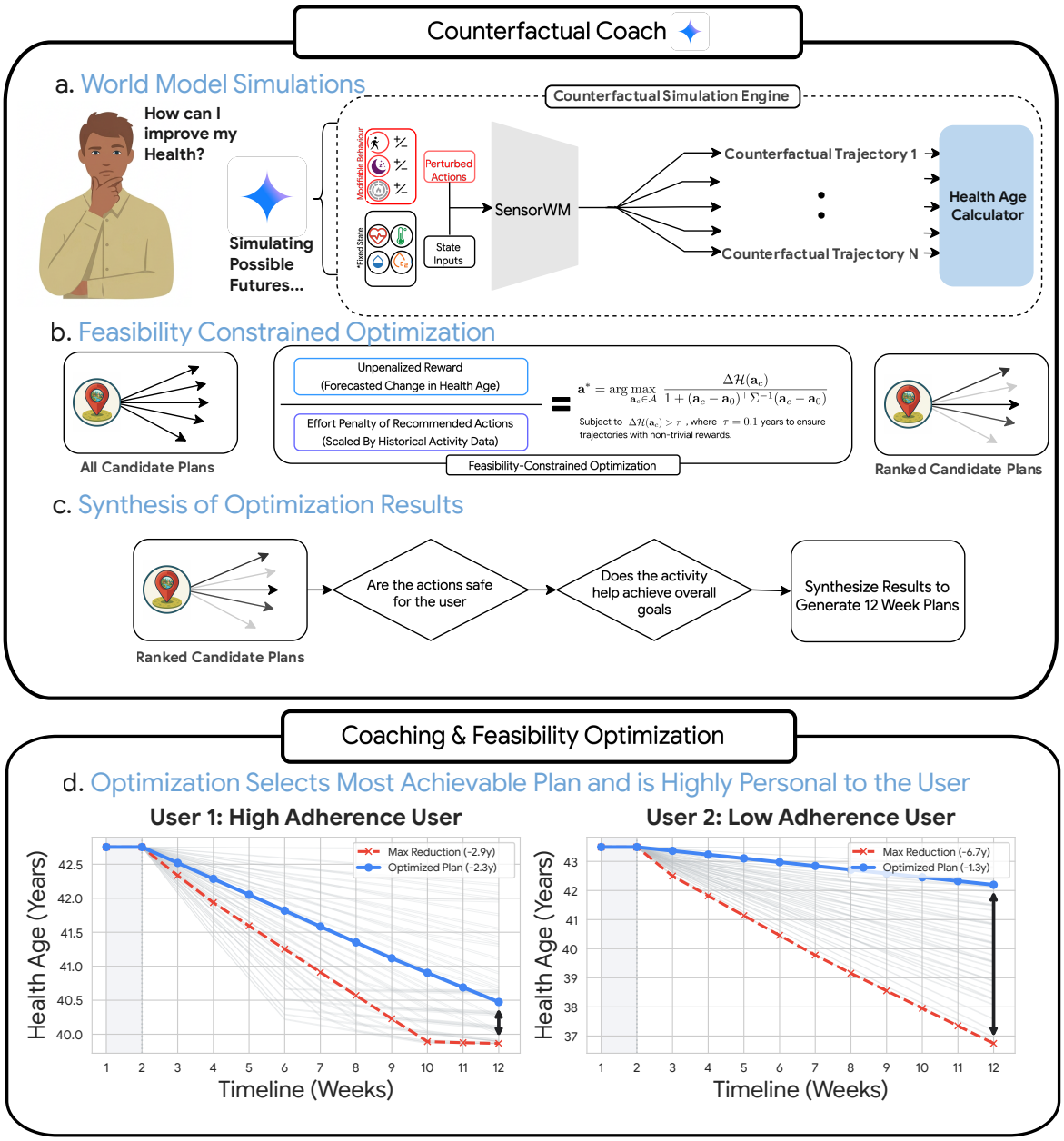


Figure 4 | **SensorWM Generates and Prioritizes Personalized Behavioral Plans for Health Coaching.** **(a)** Counterfactual simulation of 12-week health trajectories under alternative behavioral plans using SensorWM. **(b)** Feasibility-constrained optimization ranks candidate plans by projected Health Age reduction and individualized behavioral effort. **(c)** Ranked plans are screened for safety and behavioral feasibility before synthesis into 12-week coaching plans. **(d)** Representative trajectories illustrate how optimization balances achievable Health Age improvements with participant-specific historical behavioral patterns.

a feasibility-constrained optimization procedure (Figure 4b, Methods). For each participant, candidate behavioral plans were evaluated according to their projected reduction in CHA and the behavioral effort required to achieve them. Effort was defined relative to the participant's own recent behavioral variability, such that changes consistent with previously observed behavior were favored over larger departures from their usual range. The resulting candidate plans were ranked to prioritize projected physiological improvement while accounting for participant-specific feasibility.

The coaching engine then applied additional constraints before plan generation (Figure 4c). Only directly modifiable behaviors (daily steps, sleep duration, and cardio-intensity activity) were treated as intervention targets, whereas resting heart rate, other physiological measurements, and CHA remained projected outcomes. Candidate plans were further screened against participant-specific behavioral limits and safety constraints through feasibility-constrained optimization before the highest-ranked scenarios were passed to the coaching model. This design separates actions that a participant can take from physiological responses predicted by SensorWM.

At the individual level, the projected benefits of behavioral interventions remained dependent on physiological context and participant-specific feasibility (Figure 4d). For one representative participant with greater behavioral adherence capacity, feasibility-constrained optimization selected a plan associated with a 2.3-year reduction in Health Age, approaching the maximum projected reduction of 2.9 years. In contrast, for a participant with lower adherence capacity, the optimizer selected a more conservative plan yielding a 1.3-year reduction despite an unconstrained maximum of 6.7 years. These examples illustrate how maximizing projected physiological improvement alone may favor behavioral changes that are difficult to sustain, whereas incorporating individualized feasibility constraints yields more achievable and personalized intervention targets. The simulation and optimization framework identifies candidate behavioral pathways that balance anticipated health benefits against the effort required to achieve them. These ranked trajectories form the physiological basis of the Counterfactual Coaching system. We next evaluated whether providing these projected futures to an LLM-based coach improved the quality, feasibility, and clinical appropriateness of its recommendations relative to coaching based on observed wearable history alone.

2.6. Blinded human evaluation of counterfactual coaching by lifestyle coaches and clinicians

Having established that SensorWM generates individualized, feasibility ranked behavioral trajectories, we evaluated whether incorporating these projections into an LLM coaching agent improved the resulting behavioral prescriptions. We compared **Counterfactual Coaching** which conditioned the coaching agent on each participant's baseline CHA, SensorWM-projected wearable trajectory, and feasibility ranked counterfactual pathways against a **Generic Coaching** baseline that used the identical underlying LLM, participant wearable history, plan structure, and behavior-change instructions, but without CHA or simulation-derived projections.

We conducted a blinded, randomized, dual-track evaluation across 50 cases stratified equally across five prespecified WEAR-ME behavioral personas (49 valid cases after quality control; Figure 5a and Methods). Each paired plan was independently reviewed by five certified lifestyle coaches (assessing overall plan quality, anticipated long-term adherence, and 10 behavioral dimensions) and five board-certified clinicians specializing in primary care or lifestyle medicine (assessing clinical safety gates, physiological accuracy, and 6 clinical dimensions), yielding 245 paired evaluations per track across the 49 valid cases (Figures 1e and 5; Supplementary Information A.3 and Supplementary Tables 26, 27).

In blinded head-to-head comparisons, lifestyle coaches selected Counterfactual Coaching as the best overall plan in 71.4% of evaluations (175/245; 95% CI, 65.5–76.7%; P = 1.49×10^{-11}; Figure 5b), compared with 28.6% for Generic Coaching (70/245; two-sided exact binomial $P = 1.49 \times 10^{-11}$). Counterfactual Coaching was preferred for anticipated long-term adherence in 71.0% of evaluations (174/245; 95% CI, 65.0–76.3%; P = 3.72×10^{-11}; Figure 1e), compared with 29.0% for Generic Coaching (71/245; $P = 3.72 \times 10^{-11}$; Figure 1e). Preference direction was uniform across all five lifestyle coaches, with majority rater consensus favoring Counterfactual Coaching in 87.8% (43/49) of cases for overall quality and 89.8% (44/49) for adherence (Supplementary Table 26). Crucially, these gains did not come at the expense of clinical safety. Clinicians judged 99.2% (243/245) of Counterfactual plans and 97.6% (239/245) of Generic plans to be within

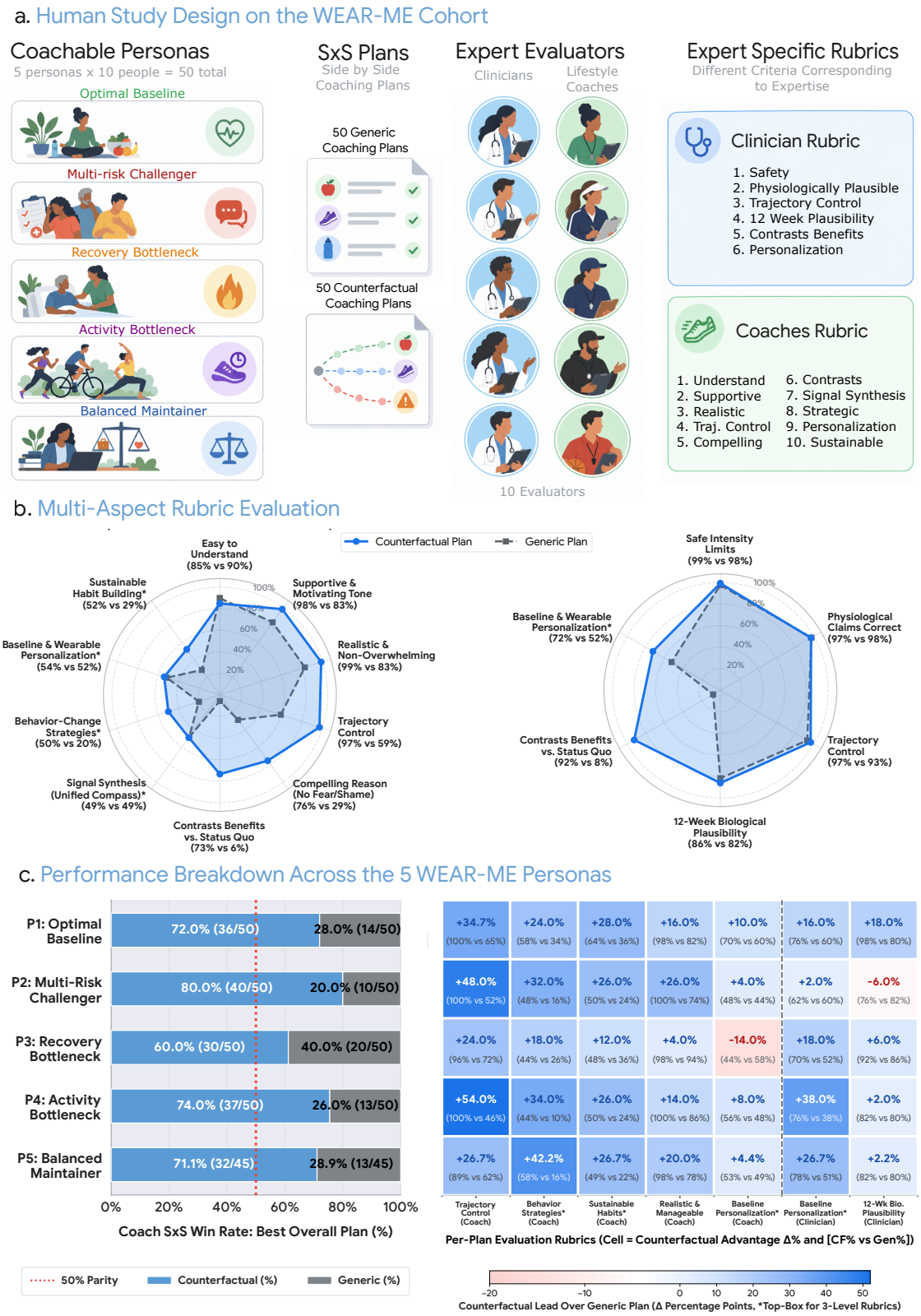


Figure 5 | **Human expert evaluation of Counterfactual vs. Generic coaching plans.** **(a)** Cohort stratification across five WEAR-ME personas and dual-track evaluation workflow (5 lifestyle coaches, 5 clinicians). **(b)** Radar comparisons of positive endorsement rates across 10 behavioral coaching rubrics (left) and 6 clinical rubrics (right). **(c)** Persona-level breakdown of coach SxS win rates (left; red dotted line indicates 50% parity) and multi-rubric percentage-point advantage (Δ = CF% – Gen%, right).

safe exercise intensity limits ($P = 0.289$), with comparably high physiological claim accuracy (97.1% versus 97.6%; Figure 1e and Supplementary Table 27).

Multi-aspect rubric evaluations revealed that overall preference was driven by dimensions directly enabled by counterfactual trajectory modeling (Figure 5b and Supplementary Tables 26, 27). Among lifestyle coaches, Counterfactual plans substantially outperformed Generic plans in contrasting the benefits of behavior change against the status-quo trajectory (73.5% versus 5.7%; $\Delta = +67.8$ percentage points, $P = 3.15 \times 10^{-45}$), providing a compelling reason to change without fear- or shame-based framing (75.5% versus 28.6%; $\Delta = +46.9$ points), and conveying actionable control over the future trajectory (97.1% versus 59.4%; $\Delta = +37.7$ points). Counterfactual plans also achieved higher top-box ratings for behavior-change strategies (50.2% versus 20.4%) and sustainable habit building (52.2% versus 28.6%), and were more frequently rated realistic and non-overwhelming (98.8% versus 82.9%) and supportive in tone (98.4% versus 83.3%), whereas general comprehensibility (84.9% versus 89.8%, $P = 0.111$) and multi-signal synthesis (49.0% versus 48.6%) were comparable between arms. Clinician evaluations mirrored these findings: clinicians rated Counterfactual plans markedly higher for contrasting intervention benefits against the status quo (92.2% versus 7.8%; $\Delta = +84.5$ points, $P = 3.49 \times 10^{-55}$), baseline physiological personalization (72.2% versus 52.2%; $\Delta = +20.0$ points, $P = 5.57 \times 10^{-6}$), and 12-week biological plausibility (86.1% versus 81.6%, $P = 0.043$). Complementary perspectives between coaches and clinicians, including how clinicians prioritized physiological risk tailoring while coaches emphasized behavioral agency and external lifestyle constraints, are detailed in Supplementary Table 28.

Stratified analyses confirmed that the preference for Counterfactual Coaching generalized across all five WEAR-ME personas (Figure 5c and Supplementary Table 29), led by *Multi-Risk Challengers* (Persona 2; 80.0% win rate, 40/50), *Activity Bottleneck* (Persona 4; 74.0%, 37/50), *Optimal Baseline* (Persona 1; 72.0%, 36/50), and *Balanced Maintainers* (Persona 5; 71.1%, 32/45). Qualitative error analysis and subgroup comparisons also surfaced two interpretable failure modes (Supplementary Table 29): first, in the *Recovery Bottleneck* persona (P3; high baseline activity but insufficient sleep, 60.0% win rate, 30/50), some plans prescribed redundant step increases in an already saturated activity domain; second, in high-risk profiles (P2), clinicians noted that a subset of 12-week CHA reductions (4–5 years) appeared overly optimistic in magnitude, indicating an opportunity for tighter physiological timescale calibration. These dual-track evaluations demonstrate that grounding an LLM coach in SensorWM counterfactual projections produced more actionable and more frequently preferred behavioral plans while maintaining high expert-rated safety across diverse user profiles.

To complement the expert evaluation and assess scalability across a larger sample, we additionally performed an automated evaluation using rubric-aligned LLM autoraters across an expanded participant cohort (Extended Data Tables 5, 6). Consistent with the blinded lifestyle-coach and clinician assessments, automated evaluations consistently favored Counterfactual Coaching over the Generic Coaching baseline across overall plan quality, trajectory control, and physiological personalization, confirming that the benefits of simulation-grounded coaching generalize across the broader cohort.

3. Discussion

In this study, we introduced CHA and embedded it within an end-to-end counterfactual coaching architecture that couples an autoregressive wearable SensorWM with a feasibility constrained coaching agent. Across population cohorts, we demonstrated that CHA, which is constructed entirely from modifiable physiological and behavioral signals passively tracked by consumer wearables, captures cardiometabolic health, providing incremental prognostic value for incident

cardiovascular and metabolic disease beyond age, sex, and BMI in the *All of Us* cohort, and within-person behavioral plasticity over time. We further showed that SensorWM accurately forecasts multi-week physiological trajectories across coupled wearable signals, outperforming generic time series foundation models and statistical baselines while generating biologically coherent counterfactual rollouts under candidate lifestyle modifications. Crucially, in a blinded, dual-track evaluation by certified lifestyle coaches and board-certified clinicians, conditioning an LLM coaching agent on CHA and SensorWM counterfactual trajectories substantially improved overall plan quality and projected long-term adherence compared with coaching grounded in retrospective wearable data alone, while maintaining high clinical safety and physiological plausibility.

A central methodological challenge in quantifying biological aging is the absence of a single gold-standard ground truth [42]. Because true biological age is latent, we evaluated CHA through its alignment with established cardiometabolic constructs, its prospective association with incident disease onset in the *All of Us* cohort, and its longitudinal sensitivity within individuals. Across academia and industry, prior efforts to quantify biological and physiological aging span four primary paradigms, each confronting a fundamental trade-off between clinical validity and everyday behavioral actionability. First, at the molecular and clinical level, epigenetic, proteomic, and blood panel clocks, evolving from first generation chronological age predictors [10, 43] to second generation mortality clocks (PhenoAge [4]; GrimAge [44, 12]), third generation longitudinal pace of aging biomarkers (DunedinPACE [45]), and clinical blood composites, [3, 46]capture deep cellular and systemic senescence, yet their reliance on invasive, costly, and episodic sampling limits their utility as high frequency feedback signals for daily behavior change. Second, classical epidemiological risk age models, such as Framingham Vascular/Heart Age [47], Spiegelhalter's effective age formulation [6], and commercial actuarial engines (Vitality Age), translate multivariable proportional hazards into intuitive age equivalents, but operate as static, episodic snapshots. Third, earlier wearable-based age estimators trained to regress chronological time [11, 5] often latch onto age correlated features that carry limited prognostic or modifiable signal, as reflected in their weak concordance with clinical reference scores in our benchmarks. Fourth, commercial consumer wearable features have largely traded multi system clinical disease grounding for convenience: Garmin Fitness Age maps cardiorespiratory fitness (VO_2max) and resting heart rate against population percentiles without linking to clinical disease incidence [48]; Oura Cardiovascular Age [49] and polysomnographic Sleep Age models [50] infer physiological age from passive photoplethysmography (pulse wave velocity) or sleep waveforms, offering observational state readouts that cannot be causally decomposed into specific behavioral targets; and WHOOP Healthspan (WHOOP Age and Pace of Aging [51]) aggregates lifestyle exposures using hazard ratios parameterized on late life all cause mortality rather than earlier onset cardiometabolic morbidity.

Our framework resolves this validity–actionability tension by unifying epidemiological survival modeling, action-conditioned trajectory forecasting, and counterfactual behavioral AI coaching agent into a modular, closed-loop architecture. Rather than regressing chronological age, matching population fitness percentiles, or awaiting late life mortality endpoints, CHA anchors high frequency wearable signals directly to published epidemiological hazard ratios for incident cardiometabolic diseases, the earliest and most modifiable burden of chronic morbidity prioritizing time varying behavioral features across physical activity, sedentary patterns, sleep regularity, and cardiorespiratory health, consistent with preventive frameworks such as Life's Essential 8 [1]. By design, this hazard ratio formulation is inherently extensible: unlike monolithic end to end machine learning clocks that require complete retraining when inputs change, new behavioral signals, wearable sensors, or clinical biomarkers (e.g., resistance training volume, dietary patterns, or continuous glucose metrics) can be seamlessly incorporated as new epidemiological evidence or sensing modalities emerge. Expressing these multi-system risks on a shared age equivalent

scale yields a responsive metric through which individuals can track their physiological state over time and transparently inspect which factors drive it. Finally, instead of treating Health Age as a static assessment, we pair CHA with an action conditioned trajectory prediction model and a counterfactual coaching agent that operates directly on the user's interpretable behavioral state space, solving for feasible, personalized habit adjustments that yield the greatest marginal reduction in cardiometabolic disease hazard and actively guide the individual toward a younger physiological trajectory.

Our work also advances recent efforts applying large language models and foundation models to biological aging and personal health coaching. Recently, Li et al. demonstrated that LLMs can estimate system-wide and organ-specific biological age directly from routine physical examination records and clinical biomarkers across large-scale populations [52]. While such approaches highlight the capacity of LLMs to synthesize cross-sectional clinical records for static risk stratification, direct zero-shot LLM estimation of health age from continuous wearable streams showed substantially lower alignment with clinical cardiometabolic references in our evaluations. Similarly, recent wearable-grounded conversational agents, including PH-LLM and multi-agent personal health frameworks [36, 35, 37] have established important benchmarks for interpreting retrospective sensor summaries, yet they rely on the language model's qualitative priors to prescribe behavioral changes. Our architecture bridges these paradigms by separating physiological state estimation and trajectory simulation from natural-language synthesis. Rather than asking an LLM to implicitly guess both biological age and future intervention responses, we pair an explicit, interpretable health age target (i.e., CHA) with a dedicated quantitative world model (i.e., SensorWM) and validate the resulting prescriptions through dual lifestyle-coaching and clinical expert tracks.

This architectural separation directly addresses the explainability gap in clinical artificial intelligence, where post-hoc attribution methods typically identify which inputs drove a static risk score without indicating how an individual can alter that trajectory [13, 53, 54]. In our framework, component-level age attributions in CHA are directly coupled to SensorWM's action-conditioned rollouts, translating static feature importance into concrete, quantified behavioral levers such as how a specific increase in daily steps, or sleep duration is projected to shift an individual's future Health Age. The impact of this design was directly reflected in our expert evaluations: the largest advantages of Counterfactual Coaching over Generic Coaching emerged precisely in contrasting projected intervention benefits against a trajectory of continued current habits (Δ = +67.8 percentage points among coaches; Δ = +84.5 points among clinicians), providing a compelling rationale for change, and fostering perceived control over one's health trajectory. Explainability, when operationalized through prospective outcome projections rather than retrospective attributions, thus serves as an active mechanism for motivational framing and sustainable habit formation.

A key technical contribution of this study is demonstrating that a domain specific Sensor World Model can learn coupled, multi week physiological dynamics across behavioral inputs and physiological responses (such as resting heart rate and physical activity), outperforming general purpose time series foundation models across forecasting horizons. Simulating "what-if" behavioral scenarios over a 12-week horizon enabled the coaching agent to evaluate candidate regimens against individualized feasibility constraints before recommending a single goal to the user.

Our dual-track human evaluation yielded important insights into both the strengths and current failure modes of simulation guided coaching, as well as how different domain experts evaluate AI-generated coaching advice. While both expert panels strongly favored Counterfactual Coaching overall and confirmed its high safety (99.2% within safe intensity bounds), lifestyle coaches and clinicians prioritized distinct dimensions of personalization. Clinicians were particularly sensitive to physiological risk tailoring and the contrast between projected intervention benefits and the continued current habits, whereas lifestyle coaches focused heavily on behavioral agency,

habit progression, and real-world barriers beyond wearable sensors noting in qualitative feedback that wearable derived personalization does not yet capture work schedules, caregiving duties, or environmental access. Furthermore, persona stratified error analyses highlighted two actionable failure modes: in already highly active users with sleep deficits (*Recovery Bottleneck*), the optimizer occasionally suggested redundant step increases alongside sleep goals; and in high-risk profiles (*Multi-Risk Challengers*), clinicians flagged a subset of 4–5 year projected CHA reductions over 12 weeks as overly optimistic, underscoring the need to constrain not only behavioral input feasibility but also the biological rate of physiological adaptation.

Our findings point toward a scalable paradigm for preventive medicine, primary care, and population health management. Because CHA and SensorWM operate entirely on passive signals from widely adopted consumer wearables, this framework can function as a continuous digital companion between episodic clinic visits identifying subclinical cardiometabolic drift early and translating longitudinal risk into individualized, clinician auditable lifestyle prescriptions. In primary care and lifestyle medicine workflows, presenting patients with an intuitive age equivalent metric alongside personalized "what-if" trajectories could enhance shared decision making, replace generic lifestyle handouts with tailored 12-week action plans, and enable remote monitoring of physiological response to prescribed habits. Prospective randomized controlled trials will be needed to determine whether deploying counterfactual wearable coaching in real-world populations leads to sustained behavioral adherence, durable reductions in CHA, and measurable improvements in long-term cardiometabolic outcomes.

Several limitations of this study warrant consideration and inform directions for future work. First, although SensorWM captures longitudinal temporal dynamics across behavioral and physiological signals, it is trained on observational wearable data rather than randomized interventional trials; consequently, its counterfactual projections reflect conditional observational associations rather than guaranteed causal treatment effects, and unmeasured time varying factors such as dietary changes, medication adjustments, intercurrent illness, or psychosocial stress may confound observed trajectories. Second, while consumer wearables enable passive, high frequency monitoring, wearable-owning cohorts can reflect socioeconomic and health conscious selection biases, and wrist worn sensors do not directly measure nutritional intake, resistance training load, or contextual barriers to behavior change such as work schedules, caregiving responsibilities, and environmental access. Third, whereas candidate sleep targets in the counterfactual grid were capped at 9 hours, we did not impose an upper saturation cap on candidate daily step targets because additional step volume can still influence projected downstream physiology such as resting heart rate and Active Zone Minutes through SensorWM; however, in already highly active individuals with insufficient sleep (*Recovery Bottleneck*), this design choice occasionally led the optimizer to propose further step increases alongside sleep goals, suggesting that future implementations should incorporate baseline-aware saturation ceilings or diminishing-return constraints in the candidate action grid. Finally, we evaluated coaching plans using blinded reviews by clinicians and lifestyle coaches, which is an essential step before deploying AI recommendations to users. A natural next step will be to follow users over time to measure how these plans affect real-world habit adherence and CHA.

4. Methods

4.1. Coachable Health Age

4.1.1. Health Age formulation

CHA was designed to summarize health on an age-equivalent scale using physiological and behavioral characteristics that can be repeatedly measured with consumer wearables and that

have established associations with adverse health outcomes. In contrast to wearable age models trained to recover chronological age, CHA was not optimized to predict elapsed time. Instead, each constituent measurement was mapped to a relative-risk contribution using published epidemiological associations, and the resulting risk contributions were expressed in age-equivalent units [6].

For participant j, the Health Age Delta was defined as

$$\Delta\text{Age}_j = \frac{1}{G}\sum_{i=1}^{K} w_i \log\left(\text{HR}_{ij}\right), \tag{1}$$

where HR_{ij} denotes the relative hazard ratio associated with health domain i for participant j, w_i denotes the corresponding weight in log-hazard space, and $G = 0.1$ is an age-rate constant approximating a 10% exponential increase in mortality risk per year of age. CHA was then

$$\text{HealthAge}_j = \text{Age}_j^{\text{chronological}} + \Delta\text{Age}_j. \tag{2}$$

Thus, $\Delta\text{Age} > 0$ denotes a Health Age older than chronological age and $\Delta\text{Age} < 0$ denotes a Health Age younger than chronological age.

The primary wearable-derived components were resting heart rate, mean daily step count, mean sleep duration, and weekly Active Zone Minutes (AZM). Cardiorespiratory fitness (VO_2 max), wearable-predicted insulin resistance, and sleep regularity were incorporated as optional components when available. Resting heart rate was mapped using published associations with incident heart-failure risk [55]; daily step count using associations with incident diabetes [56]; sleep duration using a U-shaped association with cardiovascular disease [57]; and weekly AZM using reported associations between physical-activity volume and cardiovascular risk [58, 59]. Cardiorespiratory fitness was represented relative to age-, sex-, and weight-expected VO_2 max and mapped using reported mortality associations per metabolic equivalent [60]. Sleep regularity was quantified using the standard deviation of sleep midpoint and mapped using reported associations between sleep irregularity and cardiovascular risk [61]. Predicted insulin resistance was obtained from a previously validated wearable and demographic prediction model [62].

Because the four primary wearable-derived quantities (daily steps, weekly AZM, resting heart rate, and sleep duration) encode correlated aspects of activity, autonomic tone, and sleep, summing their unadjusted literature log-hazards with unit weights would double-count shared risk variance. To account for multicollinearity among these four core domains, we estimated the empirical correlation matrix $\mathbf{R}$ of their log-hazard contributions in the WEAR-ME cohort and solved for mutually adjusted weights $\mathbf{w} = \mathbf{R}^{-1}\mathbf{1}$ in standardized log-hazard space, yielding $w_{\text{sleep}} = 0.8790$, $w_{\text{RHR}} = 0.7066$, $w_{\text{AZM}} = 0.6950$, and $w_{\text{steps}} = 0.4773$ in Eq. (1) (with $w_i = 1.0$ for the optional domains when available).

4.1.2. Detailed parameterization of CHA: core wearable domains

CHA converts each constituent wearable or physiological measurement to a log-relative-risk contribution $\log(\text{HR}_i)$ and aggregates these contributions using the domain weights w_i according to Eq. (1) [6].

4.1.2.1 Resting heart rate. Resting heart rate was mapped using reported associations with incident heart-failure risk [55], with RHR ≤ 60 bpm as the reference region (HR = 1.0):

$$\log(\mathrm{HR}_{\mathrm{RHR}}) = \begin{cases} 0, & \mathrm{RHR} \le 60, \\ \dfrac{\mathrm{RHR} - 60}{10} \log(1.08), & 60 < \mathrm{RHR} \le 70, \\ \log(1.08) + \dfrac{\mathrm{RHR} - 70}{20} \left[\log(1.33) - \log(1.08)\right], & \mathrm{RHR} > 70. \end{cases} \tag{3}$$

4.1.2.2 Daily steps. Daily step count was mapped using reported associations between daily steps and incident diabetes [56], with a reference of 8,000 steps/day (HR = 1.0) and an upper plateau at $\mathrm{Steps}_{\mathrm{eff}} = \min(\mathrm{Steps}, 9000)$:

$$\log(\mathrm{HR}_{\mathrm{steps}}) = \frac{\min(\mathrm{Steps}, 9000) - 8000}{1000} \log(0.88). \tag{4}$$

4.1.2.3 Sleep duration. Sleep duration was mapped using the reported U-shaped association with incident cardiovascular disease [57], with 6.0–9.0 hours/night as the optimal reference interval (HR = 1.0):

$$\log(\mathrm{HR}_{\mathrm{sleep}}) = \begin{cases} (6.0 - \mathrm{Sleep}) \log(1.20), & \mathrm{Sleep} < 6.0, \\ 0, & 6.0 \le \mathrm{Sleep} \le 9.0, \\ (\mathrm{Sleep} - 9.0) \log(1.34), & \mathrm{Sleep} > 9.0. \end{cases} \tag{5}$$

4.1.2.4 Active Zone Minutes. Weekly AZM was mapped using reported quartile associations between moderate-to-vigorous physical-activity volume and cardiovascular risk [58, 59], with < 115 minutes/week as the reference baseline (HR = 1.0) and a plateau at ≥ 500 minutes/week:

$$\log(\mathrm{HR}_{\mathrm{AZM}}) = \begin{cases} 0, & \mathrm{AZM} < 115, \\ \dfrac{\mathrm{AZM} - 115}{130} \log(0.76), & 115 \le \mathrm{AZM} < 245, \\ \log(0.76) + \dfrac{\mathrm{AZM} - 245}{173} \left[\log(0.71) - \log(0.76)\right], & 245 \le \mathrm{AZM} < 418, \\ \log(0.71) + \dfrac{\mathrm{AZM} - 418}{82} \left[\log(0.68) - \log(0.71)\right], & 418 \le \mathrm{AZM} < 500, \\ \log(0.68), & \mathrm{AZM} \ge 500. \end{cases} \tag{6}$$

4.1.3. Multicollinearity weights (w_i) calibrated in the WEAR-ME cohort

To estimate the domain weights w_i in Eq. (1) without double-counting shared physiological variance across correlated wearable domains, we first evaluated the four unadjusted log-hazard contributions $\log(\mathrm{HR}_{\mathrm{sleep}})$, $\log(\mathrm{HR}_{\mathrm{RHR}})$, $\log(\mathrm{HR}_{\mathrm{steps}})$, and $\log(\mathrm{HR}_{\mathrm{AZM}})$ across participants in the WEAR-ME cohort using Eqs. (3)–(6). Evaluating correlations in transformed log-hazard space rather than on raw wearable variables is necessary because raw sleep duration has a non-monotonic U-shaped relationship with mortality risk, and daily steps, RHR, and AZM contain saturation plateaus; in log-hazard space, all four domains are aligned on a common monotonic risk scale.

Because the published hazard ratios originate from independent studies with distinct clinical endpoints and covariate adjustments, rather than univariate regressions against a single outcome in a single cohort, applying an unstandardized covariance inversion ($\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}^2$) would weight

off-diagonal adjustments by the ratio of domain standard deviations (σ_k/σ_i), allowing domains with wide numerical ranges (such as daily steps) to over-subtract from narrower-range physiological domains (such as RHR). To discount each domain strictly according to its empirical informational overlap with the remaining domains, we standardized each log-hazard column to unit variance, computed the 4×4 Pearson correlation matrix $\mathbf{R}$ across the WEAR-ME cohort ($R_{ik} = \mathrm{Corr}(\log \mathrm{HR}_i, \log \mathrm{HR}_k)$), and solved the standardized linear system for the domain weights $\mathbf{w}$:

$$\mathbf{R}\tilde{\mathbf{w}} = \mathbf{1} \quad \Longleftrightarrow \quad \tilde{\mathbf{w}} = \mathbf{R}^{-1}\mathbf{1}, \qquad w_i = \min(\max(\tilde{w}_i, 0),\ 1)\,. \tag{7}$$

In Eq. (7), each domain's unit marginal weight (1.0) is decomposed into its unique direct weight w_i plus its correlated overlap with the other domains ($\sum_{k \neq i} R_{ik} w_k$). Estimating Eq. (7) on the WEAR-ME cohort yielded the weights:

$$w_{\mathrm{sleep}} = 0.8790, \qquad w_{\mathrm{RHR}} = 0.7066, \qquad w_{\mathrm{AZM}} = 0.6950, \qquad w_{\mathrm{steps}} = 0.4773. \tag{8}$$

Multiplying each core domain's log-hazard contribution by w_i in Eq. (1) scales each knot log-hazard coefficient from $\beta_i = \log(\mathrm{HR}_i)$ to $\beta_i^{\mathrm{adj}} = w_i \log(\mathrm{HR}_i) = \log(\mathrm{HR}_i^{w_i})$, which is equivalent to replacing each unadjusted knot hazard ratio HR_i in Eqs. (3)–(6) with its collinearity-adjusted hazard ratio:

$$\log\left(\mathrm{HR}_i^{\mathrm{adj}}\right) = w_i \log(\mathrm{HR}_i) \quad \Longleftrightarrow \quad \mathrm{HR}_i^{\mathrm{adj}} = \exp(w_i \log \mathrm{HR}_i) = (\mathrm{HR}_i)^{w_i}\,. \tag{9}$$

This formulation guarantees that all reference baselines remain invariant ($1.0^{w_i} = 1.0$, $\log \mathrm{HR}_i^{\mathrm{adj}} = 0$) and preserves exact piecewise continuity at all interior knots. Supplementary Table 1 summarizes the unadjusted literature hazard ratios, the WEAR-ME weights w_i, and the resulting collinearity-adjusted knot hazard ratios $\mathrm{HR}_i^{\mathrm{adj}}$.

4.1.4. Optional CHA domains (cardiorespiratory fitness, sleep regularity, and insulin resistance)

When available for a participant, the following three domains are evaluated with weight $w_i = 1.0$ (defaulting to $w_i = 0$ when any required input is unmeasured):

4.1.4.1 Cardiorespiratory fitness. When measured run $\mathrm{VO_2}$ max ($\mathrm{mL \cdot kg^{-1} \cdot min^{-1}}$), biological sex (Female = 1 for female, 0 for male), and body weight ($\mathrm{Weight_{lb}}$, in pounds) are available, expected fitness is computed via the FRIEND registry equation [60]:

$$\widehat{\mathrm{VO_2\ max}} = 79.9 - 0.39\,\mathrm{Age} - 13.7\,\mathrm{Female} - 0.127\,\mathrm{Weight_{lb}}, \tag{10}$$

and mapped to a log-hazard contribution:

$$\log(\mathrm{HR_{VO_2}}) = \left(\frac{\mathrm{VO_{2,measured}} - \widehat{\mathrm{VO_2\ max}}}{3.5}\right) \log(0.88). \tag{11}$$

4.1.4.2 Sleep regularity. When the standard deviation of sleep midpoint ($\mathrm{SD_{midpoint}}$, in minutes) is available, it is capped at $\mathrm{SD_{eff}} = \min(\mathrm{SD_{midpoint}}, 120)$ and mapped [61] with ≤ 30 minutes

as the reference region:

$$\log(\mathrm{HR}_{\mathrm{regularity}}) = \begin{cases} 0, & \mathrm{SD}_{\mathrm{eff}} \le 30, \\ \dfrac{\mathrm{SD}_{\mathrm{eff}} - 30}{30} \log(1.16), & 30 < \mathrm{SD}_{\mathrm{eff}} \le 60, \\ \log(1.16) + \dfrac{\mathrm{SD}_{\mathrm{eff}} - 60}{30} \left[\log(1.52) - \log(1.16)\right], & 60 < \mathrm{SD}_{\mathrm{eff}} \le 90, \\ \log(1.52) + \dfrac{\mathrm{SD}_{\mathrm{eff}} - 90}{30} \left[\log(2.11) - \log(1.52)\right], & \mathrm{SD}_{\mathrm{eff}} > 90. \end{cases} \tag{12}$$

4.1.4.3 Insulin resistance. When wearable-predicted insulin-resistance status [62] is available, it contributes

$$\log(\mathrm{HR}_{\mathrm{IR}}) = \begin{cases} \log(2.00), & \mathrm{IR} = \mathrm{True}, \\ 0, & \mathrm{IR} = \mathrm{False}. \end{cases} \tag{13}$$

4.1.5. Mathematical saturation and CHA plateau

Because each core physiological metric and health behavior within our framework incorporates discrete plateaus, right-censored saturation thresholds, or bounded reference ranges, the resulting CHA models naturally exhibit structural ceilings on maximum age reduction. In the unadjusted 4-metric CHA model, an individual can only achieve the maximum possible age reduction—capped at 5.13 years younger than chronological age (ΔAge = −5.135 years)—if they simultaneously optimize all four digital inputs. This requires maintaining a resting heart rate (RHR) ≤ 60 bpm (reference risk level, HR = 1.00, 0.00 years), reaching or exceeding 9,000 daily steps (capping the step HR at 0.88, or −1.28 years relative to the 8,000-step baseline), sleeping between 6.0 and 9.0 hours per night (avoiding U-shaped short- and long-sleep penalties, HR = 1.00, 0.00 years), and logging ≥ 500 weekly Active Zone Minutes (AZM; capping the AZM HR at 0.68, or −3.86 years). Because further behavioral extensions (e.g., 20,000 steps/day or 1,000 AZM/week) encounter explicit log-hazard plateaus rooted in epidemiological saturation points, the formulation mathematically prevents an individual from bypassing this lifestyle-driven rejuvenation floor through extreme volume in a single behavior.

In the collinearity-corrected CHA model, attenuating the marginal hazard ratios to account for physiological and behavioral overlap across the four core wearable metrics tightens this structural rejuvenation cap from −5.13 years to −3.29 years (ΔAge = −3.292 years). Under collinearity correction, RHR (≤ 60 bpm) and sleep duration (6.0–9.0 hours) remain neutral reference anchors (0.00 years), while the maximum achievable reductions from daily steps (≥ 9,000 steps/day, $\mathrm{HR}_{\mathrm{corr}} = 0.9408$) and weekly AZM (≥ 500 mins/week, $\mathrm{HR}_{\mathrm{corr}} = 0.7648$) attenuate to −0.61 years and −2.68 years, respectively. When the extended optional metrics are incorporated into the collinearity-corrected model, sleep midpoint consistency (≤ 30 min SD) and insulin resistance status (absent) operate strictly as unidirectional risk penalties (0.00 years when optimal, adding up to +7.47 years at ≥ 120 min SD and +6.93 years when insulin resistant) and therefore do not expand the maximum age reduction. Only the optional cardiorespiratory fitness term (VO_2 max) is modeled continuously without an upper saturation plateau, conferring an additional −1.28 years of age reduction (HR = 0.88) for each 1.0 MET (3.5 $\mathrm{mL} \cdot \mathrm{kg}^{-1} \cdot \mathrm{min}^{-1}$) achieved above age-, sex-, and weight-predicted norms, subject only to a global lower bound of 18.0 years on final Health Age.

4.1.6. Standard health and risk scores

We evaluated whether CHA captured clinically meaningful variation using two complementary reference measures of cardiovascular and metabolic health: a modified Life's Essential 8 score and the Metabolic Syndrome Severity Score. Neither reference measure was used to fit CHA.

4.1.6.1 Modified Life's Essential 8. We calculated a modified version of the American Heart Association Life's Essential 8 cardiovascular-health score [1]. Standard Life's Essential 8 integrates diet, physical activity, nicotine exposure, sleep, body mass index, blood lipids, blood glucose, and blood pressure. Because standardized dietary-intake information was unavailable, we omitted the diet component and calculated a seven-component score, hereafter referred to as mod-LE8.

Physical activity was quantified from wearable-derived AZM; sleep health from wearable-measured sleep duration; nicotine exposure from self-report; BMI from clinical measurements; blood pressure from clinical systolic and diastolic measurements together with treatment status; blood lipids from non-HDL cholesterol and lipid-lowering treatment status; and blood glucose from HbA1c and diabetes history. Each component was mapped to a 0–100 cardiovascular-health score according to the corresponding Life's Essential 8 framework [1], and mod-LE8 was calculated as the unweighted arithmetic mean of the seven available components. Higher values indicate more favorable cardiovascular health. Additional scoring details are provided in Section 4.1.12.

4.1.6.2 Metabolic Syndrome Severity Score. We quantified metabolic risk using the continuous Metabolic Syndrome Severity Score (MSSS), following the continuous latent-variable framework established using confirmatory factor analysis [63]. MSSS combines body mass index, systolic blood pressure, fasting plasma glucose, triglycerides, and high-density lipoprotein cholesterol. Scores were standardized as sex- and race/ethnicity-specific z-scores, with higher values indicating greater metabolic-syndrome severity. Participants missing one or more of the five required clinical measurements were excluded from the MSSS analysis.

4.1.7. Construct alignment, comparator benchmarks, and longitudinal modifiability

To evaluate whether CHA reflects current cardiometabolic health more faithfully than alternative age formulations, we benchmarked CHA against chronological age, blood-based Phenotypic Age (PhenoAge), a wearable chronological-age prediction model, and direct LLM-derived health-age estimates. PhenoAge was calculated according to the mortality-risk formulation of Levine et al. [4]. The wearable chronological-age comparator used embeddings from a pretrained Sensor Foundation Model (SensorFM) [20], followed by chronological-age regression and correction for the age-dependent regression-to-the-mean bias characteristic of age-prediction models [64]. Direct LLM health-age comparators were generated by prompting the base language model with either the curated CHA feature set or the full set of wearable features. The comparator implementations are described in Section 4.1.13.

Construct alignment with established cardiometabolic reference scores was assessed by calculating Pearson correlation coefficients between each age measure and modified Life's Essential 8 (mod-LE8) and Metabolic Syndrome Severity Score (MSSS) in the WEAR-ME cohort. Because higher mod-LE8 scores indicate more favorable cardiovascular health whereas higher MSSS scores indicate greater metabolic risk, absolute correlation magnitudes ($|r|$) were used to compare alignment across age formulations (Supplementary Figure 1). We additionally evaluated whether ΔCHA (Health Age minus chronological age) varied systematically across participant-reported overall health categories (from excellent to poor) by computing category-specific means and 95% confidence intervals, comparing separation against age deltas from the wearable chronological-age model.

To characterize within-person longitudinal modifiability, CHA was computed across eight consecutive three-month windows (Q1–Q8; 24 months total) among $N = 1{,}464$ participants in the longitudinal Fitbit cohort (Figure 2c). For each participant j and quarterly window $t \in \{1, \ldots, 8\}$, we computed the health-age delta as

$$\Delta\mathrm{CHA}_{j,t} = \mathrm{CHA}_{j,t} - \mathrm{Age}_{j,t}^{\mathrm{chronological}}, \tag{14}$$

and quantified longitudinal shifts relative to baseline ($\Delta\mathrm{CHA}_{j,t} - \Delta\mathrm{CHA}_{j,1}$), stratifying participants into improved ($\Delta \leq -1$ year), stable ($|\Delta| < 1$ year), and deteriorated ($\Delta \geq +1$ year) trajectory groups. These observational analyses characterize natural within-participant plasticity in CHA over coaching-relevant timescales and are not interpreted as causal effects of isolated behavioral interventions.

4.1.8. Prospective validation in the All of Us Research Program

We independently evaluated the prospective association and prognostic utility of CHA for incident cardiometabolic disease using the *All of Us* Research Program. Participants were required to have at least 730 days of Fitbit tracking span in the daily activity-summary table and at least one valid 90-day wearable window, yielding a total wearable cohort of $N = 22{,}748$ participants (276,378 three-month windows; 68,101.4 person-years of active wearable tracking; Table 1). Participant sex and year of birth were obtained from the OMOP `person` table. Baseline BMI was obtained from the OMOP `measurement` table (concept ID 3038553, restricted to physiologically plausible values of 12–80 kg/m^2) by selecting the measurement closest to the start of wearable tracking, prioritizing measurements recorded on or before the end of the 90-day baseline window. Cardiometabolic outcomes were identified from OMOP `condition_occurrence` records using the OMOP `concept_ancestor` hierarchy. Hypertension was defined using ancestor concept ID 316866, type 2 diabetes using ancestor concept ID 201820, and cardiovascular disease using ancestor concept IDs 316139, 4329847, 4182210, and 381591. The composite cardiometabolic outcome was defined as the earliest recorded diagnosis of any of these three conditions. Participants with a recorded diagnosis of hypertension, type 2 diabetes, or cardiovascular disease on or prior to the end of their baseline wearable window were excluded from incident-disease survival analyses, leaving 18,423 disease-free participants in the primary incident-risk cohort.

4.1.8.1 Wearable aggregation and data-quality criteria. Daily step count, fairly active minutes, and very active minutes were obtained from Fitbit activity-summary records. Main-sleep duration was derived from daily sleep-summary records restricted to the primary sleep episode (`is_main_sleep = TRUE`). Minute-level heart-rate records were aggregated to derive daily minimum heart rate (used as the resting heart rate proxy) and the 90th percentile of daily heart rate. A seven-day period was considered valid when at least two days contained complete activity, sleep, and heart-rate measurements. Consecutive 90-day (one-quarter) analysis windows were retained only when they contained at least five distinct valid weeks. Within each retained 90-day window, we calculated mean daily steps, mean main-sleep duration, mean daily minimum heart rate, and mean weekly active zone minutes (defined as fairly active minutes plus twice the very active minutes). CHA and Health Age Delta (ΔAge = Health Age − Chronological Age) were calculated for each 90-day window.

4.1.9. Baseline Health Age by incident disease status and disease-free survival

We first compared baseline Health Age Delta (mean ± 95% CI) between participants who remained free of all three conditions throughout follow-up (healthy controls) and those who subsequently developed incident hypertension, incident cardiovascular disease, or incident type

2 diabetes after baseline. Next, we tested whether baseline Health Age Delta stratified longitudinal cardiometabolic disease-free survival across the primary incident-risk cohort (N = 18,423). Participants were stratified at baseline into Slower/Normal Agers ($\Delta\text{Age} \leq 0$; n = 10,697) and Accelerated Agers ($\Delta\text{Age} > 0$; n = 7,726). Kaplan–Meier estimators were used to estimate the probability of remaining free of the composite cardiometabolic outcome from the end of the baseline window until the earliest recorded diagnosis of hypertension, type 2 diabetes, or cardiovascular disease, or censoring at the end of available wearable follow-up. Disease-free survival probabilities and 95% confidence intervals were evaluated at 4, 8, 16, 24, and 32 quarters (corresponding to 1, 2, 4, 6, and 8 years of follow-up; Extended Data Table 1).

4.1.10. Time-varying Cox proportional hazards analysis

To evaluate whether longitudinal Health Age Delta remained independently associated with incident cardiometabolic disease after accounting for demographic and clinical confounders, we fitted a time-varying Cox proportional hazards model. To preserve temporal ordering and mitigate reverse causation from subclinical disease during the same interval, Health Age Delta was lagged by one 90-day window so that $\Delta\text{Age}_{i,t-1}$ predicted incident diagnosis in interval t:

$$h_i(t) = h_0(t) \exp\left(\beta_\Delta \Delta\text{Age}_{i,t-1} + \beta_A \text{Age}_{i,0} + \beta_B \text{BMI}_{i,0} + \boldsymbol{\beta}_S^\top \mathbf{S}_i\right), \tag{15}$$

where $\text{Age}_{i,0}$ denotes chronological age at baseline, $\text{BMI}_{i,0}$ denotes baseline BMI, and $\mathbf{S}_i$ denotes indicator variables for sex. Participant trajectories were represented as start–stop intervals in days since baseline, with the stop time of event windows truncated at the exact diagnosis date. An L_2 ridge penalty of 10^{-4} was applied for numerical stability. The lagged time-varying analysis included 13,842 participants (164,336 intervals; 906 events); full model estimates are reported in Extended Data Table 2.

4.1.11. Incremental prognostic discrimination (Harrell's C-index)

To assess whether CHA adds genuine prognostic discrimination beyond routine clinical chart variables, we compared nested baseline Cox proportional hazards models in the disease-free cohort with complete baseline covariates (N = 14,663 participants; 1,402 incident composite diagnoses): (1) a **Baseline Clinical Model** comprising chronological age, sex, and baseline BMI alone, and (2) an extended **Baseline + CHA** model incorporating baseline Health Age Delta. Model discrimination was quantified using Harrell's concordance index (C-index). We evaluated the incremental gain in concordance ($\Delta C = C_{\text{full}} - C_{\text{base}}$) using 500-iteration paired bootstrap resampling to derive 95% confidence intervals and two-sided p-values, alongside a nested likelihood ratio test (χ^2 with 1 degree of freedom). To verify out-of-sample generalizability and rule out overfitting, we computed out-of-sample risk predictions using 5-fold event-stratified cross-validation. Finally, because chronological age alone strongly separates risk across wide age spans, we performed pre-specified age-stratified subgroup evaluations in younger and middle-aged adults (< 50 years; N = 7,399, 549 events) and older adults (≥ 50 years; N = 7,264, 853 events) to test whether CHA improves risk ranking among peers of similar chronological age (Extended Data Table 3).

4.1.12. Detailed clinical reference measures

4.1.12.1 Modified Life's Essential 8. The modified Life's Essential 8 measure was based on the American Heart Association Life's Essential 8 cardiovascular-health framework [1]. Because standardized dietary-intake data were unavailable, the dietary domain was omitted and the remaining seven cardiovascular-health domains were retained. Each domain was scored from 0

to 100 according to the corresponding Life's Essential 8 criteria, and the final mod-LE8 score was calculated as

$$\text{mod-LE8} = \frac{1}{7}\sum_{k=1}^{7} S_k, \tag{16}$$

where S_k denotes the score for each retained cardiovascular-health domain.

Physical activity was calculated from wearable-derived weekly AZM. Participants accumulating at least 150 active minutes per week received 100 points, with lower activity scaled toward zero.

Sleep health was calculated from mean wearable-measured main-sleep duration. Sleep durations between 7 and 9 hours received 100 points. Very short sleep (< 4 hours) or very long sleep (> 10 hours) received zero points, with intermediate durations scored according to the corresponding cardiovascular-health scoring scheme.

Nicotine exposure was derived from self-report. Current nicotine use or regular environmental exposure received zero points, whereas participants without reported exposure received 100 points.

BMI was calculated from clinical height and weight measurements. BMI below 25 $\mathrm{kg\,m^{-2}}$ received 100 points, with progressively lower scores at higher BMI and zero points at BMI $\geq$ 40 $\mathrm{kg\,m^{-2}}$.

Blood pressure scoring used clinical systolic and diastolic blood-pressure measurements and antihypertensive-treatment status. Untreated blood pressure below 120/80 mmHg received 100 points. Blood pressure controlled through antihypertensive treatment was capped at 80 points.

Blood-lipid scoring used non-HDL cholesterol and lipid-lowering-treatment status. Untreated non-HDL cholesterol below 130 mg/dL received 100 points, while scores for participants receiving lipid-lowering therapy were capped at 80 points.

Blood-glucose scoring was based on HbA1c together with diabetes history. Among participants without diabetes, HbA1c below 5.7% received 100 points. Among participants with diabetes, scoring reflected degree of glycaemic control, with well-controlled diabetes (HbA1c < 7.0%) receiving at most 40 points and poorly controlled diabetes (HbA1c > 9.0%) receiving 10 points.

4.1.12.2 Metabolic Syndrome Severity Score. MSSS was calculated according to the continuous confirmatory-factor-analysis framework developed for Metabolic Syndrome Severity [63]. The score incorporates BMI, systolic blood pressure, fasting plasma glucose, triglycerides, and HDL cholesterol. Scores were standardized as z-scores within sex and race/ethnicity strata, including Non-Hispanic Black, Hispanic, and Non-Hispanic White/general strata. Positive scores represent metabolic severity above the corresponding U.S. adult reference distribution. Only participants with complete measurements for all five components were included.

4.1.13. Comparator age formulations

4.1.13.1 Phenotypic Age. Phenotypic Age (PhenoAge) was calculated according to the multisystem mortality-risk formulation developed by Levine et al. [4]. The formulation uses chronological age and nine clinical biomarkers: albumin, creatinine, glucose, C-reactive protein, lymphocyte percentage, mean cell volume, red-cell distribution width, alkaline phosphatase, and white-blood-cell count. Participants missing any required variable were excluded.

The linear predictor was

$$\begin{aligned} xb = & -19.907 - 0.0336\,\text{Albumin} + 0.0095\,\text{Creatinine} + 0.1953\,\text{Glucose} \\ & + 0.0954\log(\text{CRP}) - 0.0120\,\text{Lymphocyte} + 0.0268\,\text{MCV} + 0.3306\,\text{RDW} \\ & + 0.0019\,\text{ALP} + 0.0554\,\text{WBC} + 0.0804\,\text{Age}. \end{aligned} \tag{17}$$

The corresponding cumulative hazard was

$$H(t) = \frac{\exp(xb)\,[\exp(\gamma t) - 1]}{\gamma}, \qquad t = 120, \qquad \gamma = 0.0076927, \tag{18}$$

with 10-year mortality probability

$$M = 1 - \exp[-H(t)]. \tag{19}$$

This probability was transformed to a clinical Phenotypic Age,

$$\text{ClinicalPhenoAge} = 141.50225 + \frac{\log[-0.00553\log(1 - M)]}{0.090165}, \tag{20}$$

followed by the scaling transformation used in the analysis,

$$\text{PhenoAge} = \frac{\text{ClinicalPhenoAge}}{1 + 1.28047\exp[0.0344329(\text{ClinicalPhenoAge} - 182.344)]}. \tag{21}$$

4.1.13.2 Wearable chronological-age model. As a wearable-based chronological-age comparator, daily multimodal wearable measurements were encoded using the pretrained Sensor Foundation Model (SensorFM) [20]. Participants were required to have at least 14 valid wearable days. Daily embeddings were median-pooled across valid days to obtain one participant-level embedding.

Embeddings were standardized to zero mean and unit variance and reduced using principal component analysis. The minimum number of principal components explaining at least 95% of cumulative variance was retained. An XGBoost regressor was then trained to predict chronological age using five-fold cross-validation. Hyperparameters, including L1 and L2 regularization, row and column subsampling, and tree depth, were selected within the training folds. Out-of-fold predictions were retained for evaluation.

Chronological-age prediction from physiological measurements can exhibit age-dependent estimation bias resulting from regression toward the mean [64]. We therefore fitted a linear correction between chronological age and the raw age residual. Let

$$\Delta_{\text{raw}} = \text{Age}_{\text{chrono}} - \text{Age}_{\text{pred,raw}}. \tag{22}$$

We fitted

$$\text{Bias}(\text{Age}_{\text{chrono}}) = \alpha\text{Age}_{\text{chrono}} + \beta, \tag{23}$$

and calculated

$$\Delta_{\text{unbiased}} = \Delta_{\text{raw}} - \text{Bias}(\text{Age}_{\text{chrono}}). \tag{24}$$

The bias-adjusted wearable age was then

$$\text{Age}_{\text{wearable}} = \text{Age}_{\text{chrono}} - \Delta_{\text{unbiased}}. \tag{25}$$

4.1.13.3 Exploratory LLM-derived Health Age. We additionally evaluated an exploratory LLM-derived Health Age estimate using `gemini-3.1-pro-preview`. Two wearable-input configurations were considered. The first contained multi-day WEAR-ME measurements including blood-oxygen saturation, sleep score, AZM, sleep duration, steps, and heart-rate variability. The second contained the same curated measurements used by CHA. The model was instructed to return a single integer representing estimated health age.

Generation was evaluated at temperatures $T \in \{0.1, 0.5, 1.0\}$, with three independent generations per participant and temperature. Server-side caching was disabled. Within-temperature reliability was quantified using the average participant-level standard deviation, ICC(3,1), and ICC(3,k) with $k = 3$.

Across-temperature sensitivity was assessed among participants with complete estimates at all three temperatures. Mean estimates across the three replicates were compared using the Friedman test. When the omnibus test was significant, pairwise comparisons were conducted with Wilcoxon signed-rank tests and Bonferroni correction across the three temperature comparisons.

4.2. Trajectory prediction with SensorWM

4.2.1. Pretraining data and longitudinal wearable representation

World models learn transition functions that describe how an observed system evolves over time and can be recursively applied to generate future state trajectories [33, 34, 65]. We adapted this formulation to longitudinal wearable data by representing each participant as a sequence of physiological states and observable behavioral actions and learning how physiological state evolves as a function of both recent trajectory and behavioral context.

SensorWM was developed using a large longitudinal dataset of naturalistic wearable measurements collected during routine Fitbit use. The pretraining cohort contained 96,957 participants (2.28 million total person-weeks; 182,797 valid 12-week windows). Participants contributing to the pretraining dataset provided informed consent authorizing use of de-identified data for research and health-and-wellness product development. Data collection and analysis were conducted under centralized institutional review board protocol WCG 20253840. Demographic and longitudinal characteristics of the pretraining population are reported in Table 1.

Wearable measurements spanned cardiovascular, cardiopulmonary, sleep, skin-temperature, physical-activity, and physiological-variability domains. We partitioned these measurements into physiological *state variables* and behavioral *action variables*. State variables represented the participant's observed physiological condition and included resting heart rate, heart-rate variability, respiratory rate, blood-oxygen saturation, sleep architecture, skin temperature, heart-rate-zone measurements, and longitudinal variability features. Demographic characteristics were included as static state information. Three directly observable behaviors were designated as actions: daily step count, main-sleep duration, and cardio-intensity minutes. The complete feature inventory and state–action assignments are reported in Supplementary Table 6.

Wearable measurements were harmonized to a weekly temporal resolution. Weekly representa-

tions provide a compromise between short-timescale sensor variability and the loss of temporal information associated with much coarser aggregation, and are consistent with previous approaches for representing longitudinal event streams and repeated health measurements [66, 67, 68]. This timescale also corresponds to the timescale over which many sleep and physical-activity exposures are commonly summarized [69, 70, 71].

Continuous physiological measurements, behavioral and count-based measurements were aggregated over available observation days and normalized by the number of valid days. For feature x observed on the set of valid days $\mathcal{D}_t$ during week t, the weekly aggregate was

$$x_t^{\text{week}} = \frac{\sum_{d \in \mathcal{D}_t} x_d}{|\mathcal{D}_t|}. \tag{26}$$

This normalization prevented incomplete device wear from being interpreted mechanically as reduced activity or physiological function. Continuous features were standardized using the corresponding mean and standard deviation estimated from the pretraining data,

$$\widetilde{x} = \frac{x - \mu_{\text{train}}}{\sigma_{\text{train}}}. \tag{27}$$

Each participant was subsequently represented as a longitudinal state–action sequence

$$(s_1, a_1, s_2, a_2, \ldots, s_T, a_T), \qquad T \leq 24, \tag{28}$$

where s_t denotes the physiological state during week t and a_t denotes the corresponding behavioral action vector.

4.2.2. WEAR-ME evaluation cohort

SensorWM was evaluated in an independent subset of 1,654 participants from the WEAR-ME study who contained classification labels [62]; this subset was defined by the availability of longitudinal wearable measurements alongside clinical, behavioral, medication, or laboratory-derived labels, which were not present for all 2,780 cohort participants. Evaluation participants were held out from world-model pretraining. Participants contributed an average of 19.6 ± 4.2 weeks (0.38 ± 0.08 years; up to 24 weeks) of wearable observations, providing longitudinal context for evaluating whether SensorWM preserved physiologically meaningful temporal information.

The evaluation benchmark included diabetes, hyperlipidemia, hypertension, kidney disease, respiratory disease, HOMA-IR defined insulin resistance, smoking status, menstrual regularity, and cardiovascular, diabetes, thyroid, mental-health, and general medication-use outcomes. Labels originated from participant questionnaires, medication records, or laboratory measurements depending on the task. All training, validation, and test partitions were defined at the participant level to prevent observations from the same individual appearing across data splits. Complete task definitions, sample sizes, outcome prevalence, demographic characteristics, longitudinal coverage, and task-specific split statistics are reported in Supplementary Tables 7, 9, and 10.

State–action feature specification

The complete set of state and action variables supplied to SensorWM is reported in Supplementary Table 6. The feature representation combined measurements across demographic, cardiovascular, cardiopulmonary, sleep, skin-surface, motion, and regularity or variability domains. Longitudinal wearable and event-stream representations spanning multiple physiological domains have similarly been used to capture temporally evolving health information in prior work [66, 67, 68].

Only step count, main-sleep asleep minutes, and cardio-intensity minutes were designated as behavioral actions. All remaining wearable measurements were treated as state variables, including physiological responses associated with behaviour, such as heart-rate-zone occupancy. Demographic variables were treated as static contextual state information.

This operational distinction was maintained during both pretraining and alternative-behaviour rollout. Action variables could be externally replaced by prespecified future sequences, whereas state variables were generated through the learned physiological dynamics. The state–action terminology denotes the role of a variable in SensorWM and does not imply that observational associations between actions and subsequent states identify causal intervention effects.

Temporal aggregation and preprocessing

The underlying Fitbit measurements were available at heterogeneous temporal resolutions ranging from minute-level physiological streams to daily summary measurements. Directly modeling transitions at the highest available sampling frequency would emphasize substantial short-timescale variation and increase sequence length, whereas very coarse aggregation would reduce the number of available longitudinal transitions. We therefore represented the wearable data on a common weekly grid. Similar temporal harmonization is commonly used for longitudinal health-event representations [66, 67, 68], and weekly summaries correspond naturally to established reporting intervals for sleep and physical activity [69, 70, 71].

Continuous physiological variables, including heart-rate variability, respiratory rate, blood-oxygen saturation, and temperature-related measurements, were summarized within each weekly interval. Behavioral and count-based variables were aggregated over valid observation days and normalized by the number of observed days according to Eq. 26. Continuous variables were subsequently standardized using pretraining-set statistics according to Eq. 27. The same stored normalization parameters were used during downstream inference and to invert decoded predictions to their original physiological units.

4.2.3. Action-conditioned latent dynamics

SensorWM was adapted from joint-embedding predictive approaches in which an encoder and latent transition model are optimized jointly without requiring reconstruction of the original observations [34, 40]. Let

$$s_t \in \mathbb{R}^{n_s}, \qquad a_t \in \mathbb{R}^{n_a}, \tag{29}$$

denote the physiological state and behavioral action vector, respectively, during week t. A state encoder E maps the physiological state to a $d = 128$ dimensional latent representation,

$$z_t = E(s_t), \qquad z_t \in \mathbb{R}^{128}. \tag{30}$$

SensorWM models second-order dynamics in latent physiological space,

$$p(z_{t+1} \mid z_t, z_{t-1}, a_t). \tag{31}$$

Equivalently, the process can be represented as first-order Markov in an augmented state containing the current latent representation and its recent displacement,

$$\Delta z_t = z_t - z_{t-1}. \tag{32}$$

The inclusion of Δz_t allows the model to distinguish participants occupying similar current physiological states but moving through latent space in different directions.

The encoder was implemented as a multilayer perceptron with hidden dimension 256 and latent dimension 128, using batch normalization, ReLU nonlinearities, and dropout. This latent dimensionality provides sufficient capacity for multivariate temporal representations while limiting the dimensionality of the learned transition space [72].

The latent transition model was parameterized by a predictor network g. Given the current latent state and behavioral action, the predictor generated a state-dependent linear transition operator,

$$F_t = g(z_t, a_t), \qquad F_t \in \mathbb{R}^{128\times 128}. \tag{33}$$

Thus, SensorWM does not learn a single global linear dynamical system. Instead, the transition operator varies as a function of the participant's current physiological representation and behavioral context.

Behavioral information was incorporated into the predictor using adaptive layer normalization and a multiplicative action gate following the action-conditioning formulation used in LeWorldModel [40]. The action-conditioning scale and shift parameters were initialized at zero, such that optimization began near an unconditioned transition model and action dependence was learned progressively. Spectral normalization was applied to linear layers of the predictor to constrain the magnitude of the transition mapping during recursive rollout.

Rather than predicting each future latent state independently, SensorWM transforms the participant's preceding latent displacement. The predicted displacement is

$$\widehat{\Delta z}_{t+1} = F_t \Delta z_t, \tag{34}$$

and the subsequent latent state is obtained by discrete integration,

$$\widehat{z}_{t+1} = z_t + \widehat{\Delta z}_{t+1}. \tag{35}$$

Combining Eqs. 32–35,

$$\widehat{z}_{t+1} = z_t + g(z_t, a_t)\,(z_t - z_{t-1})\,. \tag{36}$$

This parameterization represents physiological evolution as an action-dependent transformation of the participant's recent direction of movement through latent space. The learned operator can preserve, attenuate, amplify, or rotate the previous displacement according to the current physiological state and supplied behavior.

Sketched Isotropic Gaussian Regularization

SensorWM used Sketched Isotropic Gaussian Regularization (SIGReg), following the representation regularization used in LeWorldModel [40], to discourage collapse of the jointly learned latent representation. The target aggregate latent distribution was an isotropic standard Gaussian,

$$z \sim \mathcal{N}(0, I). \tag{37}$$

Rather than estimating the full high-dimensional distribution directly, SIGReg projected mini-batch latent representations onto randomly sampled one-dimensional directions and quantified

deviation from a standard Gaussian using an Epps–Pulley characteristic-function criterion. Each regularization step used 256 random projections.

SIGReg was applied jointly with the rollout and trajectory-path objectives described in the main Methods. It therefore provided an anti-collapse constraint without requiring an input-reconstruction objective, explicit contrastive negative examples, a momentum-updated target encoder, or stop-gradient target representations.

4.2.4. Multi-step trajectory training

SensorWM was trained using 12-week longitudinal rollout windows. Each trajectory was initialized from the first two observed physiological states,

$$\widehat{z}_0 = z_0^* = E(s_0), \qquad \widehat{z}_1 = z_1^* = E(s_1), \tag{38}$$

with initial latent displacement

$$\widehat{\Delta z}_1 = z_1^* - z_0^*. \tag{39}$$

Subsequent states were generated recursively. At rollout step t,

$$\widehat{F}_t = g(\widehat{z}_t, a_t), \tag{40}$$

$$\widehat{\Delta z}_{t+1} = \widehat{F}_t \widehat{\Delta z}_t, \tag{41}$$

$$\widehat{z}_{t+1} = \widehat{z}_t + \widehat{\Delta z}_{t+1}. \tag{42}$$

After initialization, predicted states rather than future observed states were supplied to subsequent transitions. Prediction errors at earlier rollout steps could therefore propagate to later steps during training, exposing the model to the same recursive dynamics subsequently used during inference.

For each future week, the latent target was produced by applying the jointly trained encoder to the observed physiological state,

$$z_{t+1}^* = E(s_{t+1}). \tag{43}$$

The primary rollout objective was

$$\mathcal{L}_{\text{roll}} = \frac{1}{|\mathcal{M}|} \sum_{t \in \mathcal{M}} \left\| \widehat{z}_{t+1} - z_{t+1}^* \right\|_2^2, \tag{44}$$

where $\mathcal{M}$ denotes the set of valid transitions in the training window.

We additionally constrained the local direction of the learned trajectory. The empirical displacement between consecutively encoded observed states was

$$\Delta z_{t+1}^* = z_{t+1}^* - z_t^*, \tag{45}$$

and the trajectory-path objective was

$$\mathcal{L}_{\text{path}} = \frac{1}{|\mathcal{M}|} \sum_{t \in \mathcal{M}} \left\| \widehat{\Delta z}_{t+1} - \Delta z^*_{t+1} \right\|_2^2 . \tag{46}$$

The rollout and path objectives therefore constrain complementary properties of the learned dynamics: $\mathcal{L}_{\text{roll}}$ penalizes errors in the predicted latent location, whereas $\mathcal{L}_{\text{path}}$ penalizes errors in the direction and magnitude of local movement. The complete SensorWM training objective was

$$\mathcal{L}_{\text{SensorWM}} = \lambda_{\text{MSE}} \mathcal{L}_{\text{roll}} + \lambda_{\text{path}} \mathcal{L}_{\text{path}} + \lambda_{\text{SIG}} \mathcal{L}_{\text{SIG}}, \tag{47}$$

with $\lambda_{\text{MSE}} = 1$. During model development, λ_{path} was evaluated over the range 1.5–3.0 dynamically adjusted during pre-training and $\lambda_{\text{SIG}} = 0.1$. The implementation of SIGReg is described in Methods 4.2.3.

The encoder and transition predictor were optimized jointly using AdamW with an initial learning rate of 3×10^{-5} and weight decay of 10^{-3}. Global gradient norm was clipped at 1.0.

4.2.5. Diagnostic physiological decoder and recursive forecasting

SensorWM pretraining was performed entirely in latent space and did not require reconstruction of the original wearable measurements, consistent with the joint-embedding predictive formulation on which the architecture was based [34, 40]. To evaluate predicted trajectories in physiological units, we subsequently trained a post hoc linear decoder D,

$$\widehat{s}_t = D(z_t), \tag{48}$$

while keeping the pretrained encoder and transition predictor fixed. The decoder was optimized using masked mean-squared error such that only observed physiological measurements contributed to the loss. AdamW was used with learning rate 10^{-3} and weight decay 10^{-4}. Fifteen percent of decoder-training observations were reserved for internal validation. Training used early stopping, and the parameters corresponding to the lowest validation loss were restored before evaluation. The decoder did not contribute gradients to the encoder or transition predictor and therefore did not alter the pretrained SensorWM representation.

For action-conditioned forecasting, each participant was initialized from the two most recent observed physiological states,

$$z_0 = E(s_0), \qquad z_1 = E(s_1), \qquad \Delta z_1 = z_1 - z_0. \tag{49}$$

SensorWM was then propagated autoregressively using the participant's recorded behavioral action sequence over the forecasting horizon. At each prediction step, the corresponding behavioral action vector was supplied as an exogenous input to condition the physiological transition. Future physiological states were predicted recursively without access to future observed physiological measurements.

$$F_t = g(z_t, a_t), \tag{50}$$

$$\Delta z_{t+1} = F_t \Delta z_t, \tag{51}$$

$$z_{t+1} = z_t + \Delta z_{t+1}. \tag{52}$$

No future physiological measurements were provided after rollout initialization. Subsequent physiological states were predicted autoregressively using the action-conditioned latent transition dynamics defined in Eq. 52.

Feature-level forecasting performance was quantified using the coefficient of determination (R^2) and Pearson correlation coefficient over the population. SensorWM was compared with the mean of the preceding one, two, and four weeks, last observation carried forward (LOCF), and TimesFM [41] as a general-purpose time-series forecasting model.

Complete feature-level R^2 and correlation results are reported in Supplementary Tables 11–12.

4.2.6. Trajectory and downstream information fidelity

We evaluated whether SensorWM-generated trajectories preserved both the geometric structure and clinically relevant information contained in observed longitudinal wearable measurements.

To visualize trajectory geometry, observed wearable sequences were passed through the frozen SensorWM encoder to obtain 128-dimensional latent representations at each weekly time point and compared against the generated latent features. Principal-component analysis was fitted to the combined set of observed and generated latent states, and the same transformation was applied to both. The first two principal components were used to visualize participant-specific trajectories, enabling direct comparison of their direction, dispersion, and temporal evolution in a common coordinate system.

We next evaluated whether generated trajectories retained information relevant to downstream health prediction. For each task, the observed longitudinal Fitbit sequence and corresponding SensorWM-generated trajectory were evaluated using matched gated recurrent unit (GRU) sequence classifiers [73]. The classifier architecture, optimization procedure, train–validation–test partition, and outcome definition were held fixed between observed and generated inputs so that performance differences reflected information retained by the generated trajectory rather than differences in downstream model capacity.

We additionally compared the trajectory-based representations with two non-trajectory baselines derived from the same wearable data: features obtained from the first week of wearable observations and embeddings generated by the pretrained SensorFM [20]. Representations were mapped to a common 128-dimensional space for controlled comparison. The observed and SensorWM-generated longitudinal sequences were evaluated with the same GRU-based sequence-classification framework [73], whereas the comparison representations supplied static participant-level information.

Classification performance was quantified using area under the receiver operating characteristic curve (AUROC). For the primary comparison between observed and SensorWM-generated trajectories, differences in paired AUROCs were assessed using DeLong's test [74] for correlated receiver operating characteristic curves.

Failure to reject the null hypothesis was interpreted as an absence of a detectable difference rather than as formal statistical equivalence. Complete classification results and additional robustness analyses are reported in Supplementary Tables 15–24 and Supplementary Results A.

4.2.7. Alternative-behavior rollouts

The same frozen transition model was subsequently evaluated under prespecified alternative behavioral sequences. For each participant, the physiological initialization was held fixed while the recorded future action sequence $\{a_t\}_{t=1}^{K}$ was replaced with an alternative sequence $\{\widetilde{a}_t\}_{t=1}^{K}$. SensorWM was then propagated autoregressively using the behavioral action vector at each

forecast step. Specifically, the transition operator was conditioned on the current predicted latent physiological state and the corresponding action vector, and was applied to the preceding latent state difference to predict the subsequent physiological state.

$$\widetilde{z}_{t+1} = \widetilde{z}_t + g(\widetilde{z}_t, \widetilde{a}_t)\,(\widetilde{z}_t - \widetilde{z}_{t-1})\,, \tag{53}$$

and decoded to physiological space as

$$\widetilde{s}_t = D(\widetilde{z}_t). \tag{54}$$

Observed-action and alternative-action trajectories therefore shared the same initial physiological history and differed in the behavioral sequence supplied to the learned transition model. Their divergence represents a model-implied conditional difference under the specified action sequences. Because SensorWM was trained using observational wearable data, these alternative trajectories were interpreted as conditional model projections rather than identified causal effects of behavioral intervention.

4.3. Agentic Counterfactual Coaching

4.3.1. WEAR-ME cohort and feature preprocessing

Participants for the coaching evaluations were drawn from the WEAR-ME study (N = 2,780 participants with complete demographic linkage and at least three valid weeks of longitudinal wearable data). Inputs to plan generation comprised wearable-derived baseline metrics, chronological age, CHA, resting heart rate (RHR), daily steps, device-measured sleep duration, Active Zone Minutes (AZM), and cardiorespiratory fitness, together with each participant's habitual distributions (up to 24 weeks; mean 19.6 ± 4.2 weeks, corresponding to 0.38 ± 0.08 years and 3,289 complete 12-week windows in Table 1), historical peaks, and model-projected trajectories. All data were de-identified and accessed under the study's institutional governance and privacy controls.

Per-domain daily wearable records were aggregated to ISO calendar weeks by computing the weekly mean of daily values, yielding one weekly feature vector per participant. Static features were harmonized to standard units (weight in kilograms, height in meters, body mass index [BMI] in kg/m^2), and chronological age was derived from birth year. Each feature was standardized to zero mean and unit variance using pre-trained cohort-level statistics. Features were partitioned into participant-controllable *actions* (volitional behaviors: daily steps, main-sleep duration in minutes, and cardiorespiratory exercise minutes) and observed physiological *states* (RHR, AZM, and remaining physiological markers). Sequences were padded to a uniform length with an accompanying validity mask.

4.3.2. Counterfactual simulation and feasibility-constrained optimization

For each participant, we simulated a set of feasible 12-week behavior-change scenarios and their projected physiological consequences, casting plan selection as a constrained optimization problem that trades projected CHA improvement against individualized behavioral effort.

Candidate space and counterfactual rollouts. By construction, candidate 0 (all targets equal to baseline) represents the "do-nothing" scenario, in which observed behavioral levels are held constant over the 12-week forecasting horizon. Its projected endpoint defines the *status-quo* anchor, $\mathrm{CHA}_{\mathrm{sq}}$. For each intervention candidate c, we defined the projected counterfactual benefit as $\Delta\mathrm{CHA}(c) = \mathrm{CHA}_{\mathrm{sq}} - \mathrm{CHA}_c$, where CHA_c denotes the corresponding 12-week projected CHA.

This future-versus-future comparison measures the incremental benefit of behavioral modification relative to continued habitual behavior, rather than the absolute change from the participant's observed baseline CHA, thereby accounting for physiological drift predicted under unchanged actions.

Personalized effort and optimization objective. To reflect individual behavioral flexibility rather than population-average difficulty, we estimated per-action variability σ_k^{user} as the within-participant standard deviation of action k over the participant's historical wearable record, bounded below by safety floors (500 steps, 15 sleep minutes, and 10 cardiorespiratory exercise minutes) so that highly consistent participants were not assigned near-infinite effort penalties. Personalized effort was defined as the sum of squared, self-standardized deviations from baseline:

$$E(c) = \sum_{k \in \{\text{step, sleep, cardio}\}} \left(\frac{c_k - b_k}{\sigma_k^{\text{user}}} \right)^2 . \tag{55}$$

Each candidate trajectory was scored by its projected CHA reduction per unit of personalized effort,

$$\text{score}(c) = \frac{\Delta\text{CHA}(c)}{1 + E(c)} = \frac{\text{CHA}_{\text{sq}} - \text{CHA}_c}{1 + E(c)}, \tag{56}$$

where the additive constant regularizes the ratio at $E(c) = 0$. We discarded candidates whose projected reduction did not exceed a minimal clinically meaningful threshold ($\text{CHA}_{\text{sq}} - \text{CHA}_c >$ 0.1 years) and ranked the remaining feasible trajectories by Eq. (56). The ranked candidates were subsequently organized into three behavioral pathways corresponding to sleep-focused, movement-focused, and combined interventions. Up to three pathways were provided to the coaching agent as Option 1, Option 2, and Option 3, together with the status-quo projection CHA_{sq}, participant-specific historical wearable ranges, and recorded behavioral peaks. Each option described the proposed behavioral targets and their projected 12-week changes in CHA and resting heart rate, allowing the agent to synthesize a personalized coaching plan from distinct behavioral alternatives. Sensitivity analyses across objective hyperparameters are reported in the Supplementary Information.

Action–outcome firewall. The optimization exposes only volitional behavioral actions (steps, sleep duration, cardiorespiratory exercise) as recommendable targets; projected physiological states (RHR, AZM, and CHA) are treated strictly as downstream model outcomes and are never prescribed as direct behavioral goals. This separation is enforced in the generation prompt and audited in evaluation, preventing the agent from instructing a user to "lower your resting heart rate" rather than prescribing the behavioral modifications projected to achieve that outcome.

4.3.3. Paired coaching-plan generation and two-arm study design

To isolate the contribution of CHA and SensorWM counterfactual projections, we used a two-arm, within-participant paired design. For each participant, two 12-week personalized lifestyle coaching plans were generated from identical baseline demographic and longitudinal wearable histories (up to 24 weeks; recent daily averages, habitual means, standard deviations, minimum–maximum ranges, and historical peaks) using the same underlying language model (`gemini-3.5-flash`, temperature 0.2, served via the OneTwo interface):

- **Generic Coaching Arm (Baseline Control):** Conditioned on the participant's demographics, baseline wearable metrics, 24-week habitual distributions, and historical peaks, but **without**

access to CHA or SensorWM counterfactual simulations. The model was instructed to synthesize the wearable history and construct a progressive 12-week behavior-change plan across three 4-week phases (*Weeks 1–4: Foundation*, *Weeks 5–8: Build*, and *Weeks 9–12: Sustain*).

- **Counterfactual Coaching Arm (Full System):** Conditioned on the identical demographic and longitudinal wearable context **plus** the participant's baseline CHA, the 12-week status-quo projection (*Future A: If No Habit Changes*, CHA_{sq}), and the top feasibility-ranked counterfactual pathways (*Future B: Your Counterfactual Options*), quantifying projected changes in CHA (ΔCHA), RHR (ΔRHR), and AZM alongside historical-best self-efficacy anchors.

To prevent confounding from stylistic or structural formatting differences, both arms were governed by strict **Universal Tone and Formatting Parity Guardrails** (Supplementary Information C.3), enforcing an identical 500–700-word target length, an identical four-part layout (`Your Health Snapshot`, `12-Week Progressive Plan`, `Evidence-Based Habit Formation`, and `Summary`), explicit Self-Determination Theory (SDT) motivational interviewing principles (implementation intentions, coping planning, self-monitoring, and autonomy support), and a prohibition against clinical jargon or fear/shame-based framing. All plan generation and evaluations were conducted offline using retrospective participant profiles.

Experimental Prompt Architectures and Universal Tone Parity Guardrails

To ensure a rigorous scientific comparison between the **Generic Coaching Arm** and the **Counterfactual Coaching Arm**, both prompts were engineered under strict **Tone, Structure, and Length Parity Guardrails**. In preliminary pilot testing, unaligned baseline prompts tended to produce dry, bulleted clinical summaries, whereas counterfactual prompts produced warm, narrative coaching, confounding the informational value of counterfactual simulations with stylistic warmth. We eliminated this confounder by enforcing identical persona instructions, identical Self-Determination Theory (SDT) motivational interviewing principles, identical formatting constraints, and an identical four-section output schema across both arms:

1. **Tone & Empathy:** Warm, supportive, non-judgmental, collaborative, and empowering; strict prohibition against fear-based, alarmist, or shame-inducing language.
2. **No Clinical or Statistical Jargon:** Explicit ban on raw statistical or variable tokens (such as `std`, `min/max`, `wearable_rhr`, `counterfactual`, `baseline delta`, or `intervention vector`). All physiological concepts (such as Resting Heart Rate and Active Zone Minutes) must be explained in accessible, everyday language.
3. **Target Length Parity:** Strictly bounded to **500–700 words** across both arms so raters could not distinguish arms by verbosity.
4. **Four-Section Structural Parity:**
    - **Section 1:** `YOUR HEALTH SNAPSHOT` (Generic) / `YOUR HEALTH SNAPSHOT & HEALTH AGE` (Counterfactual).
    - **Section 2:** `YOUR 12-WEEK PROGRESSIVE PLAN` (Generic) / `YOUR COUNTERFACTUAL PATHS & 12-WEEK PLAN` (Counterfactual), structured identically into *Phase 1 (Weeks 1–4: Foundation)*, *Phase 2 (Weeks 5–8: Build)*, and *Phase 3 (Weeks 9–12: Sustain)*.
    - **Section 3:** `EVIDENCE-BASED HABIT FORMATION` (incorporating habit stacking, implementation intentions, self-monitoring, and barrier troubleshooting).
    - **Section 4:** `SUMMARY & NEXT STEPS` (closing with an empowering one-sentence takeaway).

4.3.4. Dual-track human expert evaluation and cohort stratification

Persona stratification and sampling. To evaluate clinical safety, behavioral quality, and anticipated adherence across diverse user profiles without over-representing median wearable users, we partitioned the WEAR-ME cohort into five prespecified Coachable Personas defined by baseline daily steps, sleep duration, BMI, and ΔCHA (CHA – chronological age**, where CHA denotes the formulation unadjusted for multicollinearity**):

1. **Persona 1 — Optimal Baseline:** High physical activity (≥ 10,000 steps/day; sub-cohort mean 13,775 steps/day), adequate sleep (7–9 h/night; mean 7.42 h), healthy BMI (18.5–24.9 kg/m^2; mean 22.9 kg/m^2), and favorable ΔCHA < −1.0 year (mean −4.99 years; n = 78, 2.81% of cohort), testing maintenance coaching and ceiling effects.
2. **Persona 2 — Multi-Risk Challenger:** Sedentary activity (< 5,000 steps/day; mean 3,341 steps/day), insufficient sleep (< 7.0 h/night; mean 6.03 h), elevated BMI (≥ 25.0 kg/m^2; mean 36.3 kg/m^2), and accelerated ΔCHA ≥ +1.0 year (mean +3.06 years; n = 277, 9.96% of cohort), testing multi-lever prioritization and psychological feasibility.
3. **Persona 3 — Recovery Bottleneck:** High physical activity (≥ 10,000 steps/day; mean 12,738 steps/day) paired with chronic sleep restriction (< 7.0 h/night; mean 6.23 h; n = 354, 12.73% of cohort), testing whether the system recognizes activity saturation and isolates sleep and recovery.
4. **Persona 4 — Activity Bottleneck:** Isolated inactivity (< 5,000 steps/day; mean 3,190 steps/day) with preserved sleep duration (7–9 h/night; mean 7.67 h) and accelerated ΔCHA ≥ +1.0 year (mean +2.78 years; n = 227, 8.17% of cohort), testing targeted physical activity escalation.
5. **Persona 5 — Balanced Maintainer:** Moderate daily movement (5,000–10,000 steps/day; mean 5,539 steps/day), adequate sleep (7–9 h/night; mean 7.48 h), and age-aligned ΔCHA $\in [-1.0, +1.0]$ year (mean −0.31 years; n = 143, 5.14% of cohort), testing progressive habit building in typical users.

Ten participants were randomly sampled from each persona (N = 50 cases total). During pre-lock quality control, one case in Persona 5 (Participant ID 57217) was identified as having received identical plan text in both blinded columns due to a serialization collision and was excluded *a priori*, leaving N = 49 valid participant cases (10 each in Personas 1–4 and 9 in Persona 5; mean chronological age 49.5 ± 6.6 years, 57.1% female, mean BMI 31.2 ± 7.2 kg/m^2, mean baseline CHA 53.6 ± 6.1 years, ΔCHA = +4.04 ± 4.52 years). For each case, the Generic and Counterfactual plans were randomly assigned to `Plan 1` and `Plan 2` via a Bernoulli coin flip (49.0% versus 51.0%), with all raters blinded to study hypotheses and arm assignments.

Dual-track expert panels and evaluation rubrics. We convened two complementary panels of domain experts (N = 10 raters total):

1. **Certified Lifestyle & Health Coaches (n = 5):** Board-certified health, wellness, and exercise coaches evaluated each plan across **10 behavioral rubric dimensions**—comprising 6 binary (`Yes`/`No`) dimensions (*Easy to Understand*, *Supportive & Motivating Tone*, *Trajectory Control*, *Compelling Reason [No Fear/Shame]*, *Contrasts Benefits vs. Status Quo*, and *Realistic & Non-Overwhelming*) and 4 three-level ordinal dimensions (*Signal Synthesis*, *Baseline & Wearable Personalization*, *Behavior-Change Strategies*, and *Sustainable Habit Building*)—followed by **2 forced-choice side-by-side (SxS) primary endpoints**: **Best Overall Coaching Plan** and **Highest Likelihood of Long-Term Adherence**, with mandatory free-text rationale.
2. **Board-Certified Clinicians (n = 5):** Physicians in internal medicine, preventive medicine, and lifestyle/cardiovascular medicine independently audited each plan across **6 clinical dimensions**—comprising 5 binary clinical safety and plausibility gates (*Safe Intensity Limits*,

Physiological Claims Correct, *12-Week Biological Plausibility*, *Trajectory Control*, and *Contrasts Benefits vs. Status Quo*) and 1 three-level ordinal clinical personalization rubric (*Baseline & Wearable Personalization*).

Three shared dimensions, Contrasts Benefits vs. Status Quo, Trajectory Control, and Baseline & Wearable Personalization, used identical rubric definitions in both expert panels to enable cross-disciplinary concordance analysis. Item-specific denominators are reported in Supplementary Tables 26 and 27 with cross-disciplinary comparisons summarized in Table 28.

4.3.5. Full-cohort automated evaluation with blinded LLM autoraters

To complement the 49-case human expert study and evaluate plan quality across the full eligible WEAR-ME cohort (N = 2,780; $n = 2,613$ evaluable paired plans after quality control), we conducted a blinded automated evaluation using a pre-specified, Kirkpatrick-Model inspired four-level rubric [75] (Extended Data Table 4). Each of the 12 dimensions across Level 1 (*Reaction*: Clarity & readability, Tone & motivation, Personal relevance), Level 2 (*Learning*: Data comprehension, Causal understanding, Insight/stakes), Level 3 (*Behavior*: Actionability & specificity, Feasibility & progression, Behavior-change technique), and Level 4 (*Results*: Predicted health impact, Measurability, Evidence grounding) was scored on an anchored 1–5 scale. We computed both an *equal-weighted* overall score (25% per level) and a *priority-weighted* overall score (Reaction 0.15, Learning 0.30, Behavior 0.35, Results 0.20) [76],

$$\text{overall} = \sum_{\ell=1}^{4} w_\ell L_\ell, \qquad \sum_{\ell=1}^{4} w_\ell = 1, \tag{57}$$

alongside a four-item *safety and accuracy gate* rating **unsafe advice**, **fabricated numbers**, **missing medical disclaimer**, and **factual/physiological error** as *none*, *minor*, or *major* (with any *major* flag capping overall scores at 2.0, alongside an uncapped quality index reported separately).

For each participant, the paired Generic and Counterfactual plans were randomly shuffled and relabeled as "Plan 1" and "Plan 2" prior to judging. The autorater received a ground-truth block of the participant's observed metrics, historical peaks, and model projections, with explicit instructions to evaluate substance rather than length or vocabulary and to treat clearly framed forward-looking projections as legitimate rather than fabricated numbers. Each paired plan set was evaluated three times independently (N = 3 passes; primary judge `gemini-3.5-flash`, temperature 0.5), aggregating dimension scores by median, safety flags by worst-case severity, and head-to-head preference by majority vote across passes.

4.3.6. Statistical analysis

All primary comparisons in both the human expert study and the full-cohort automated evaluation used within-participant paired contrasts between the Counterfactual and Generic coaching arms. In the human expert evaluation (N = 49 cases; 245 paired evaluations per track, 245 coach evaluations with complete head-to-head responses in Figures 1e and 5), forced-choice head-to-head endpoints (*Best Overall Plan* and *Highest Likelihood of Long-Term Adherence*) were tested against the null hypothesis of equal preference ($p_0 = 0.50$) using two-sided exact binomial tests with Wilson 95% confidence intervals, supplemented by case-level majority consensus ($\geq$ 3 of 5 raters). For per-plan binary and 3-level ordinal rubrics, we prespecified top-box achievement (`Yes` for binary items; highest anchor for ordinal items) and tested paired percentage-point differences (Δ = Counterfactual% – Generic%) using two-sided exact McNemar's tests on discordant pairs. In the full-cohort automated evaluation, paired continuous rubric contrasts were summarized by the mean within-participant difference with a 10,000-sample percentile bootstrap 95% confidence

interval, a two-sided paired Wilcoxon signed-rank test (with Holm–Bonferroni correction across rubric families), and three paired effect sizes (Cohen's d_z, matched-pairs rank-biserial correlation, and Cliff's δ). Binary preference rates were summarized with Wilson 95% confidence intervals, and paired safety flag rates were compared using McNemar's test. Autorater self-consistency across repeated passes was quantified using Fleiss' κ and Krippendorff's α, and cross-model inter-judge agreement (`gemini-3.5-flash` versus `gemini-3.1-pro-preview`) was assessed via intra-class correlation ICC(2, 1), Spearman's ρ, quadratic-weighted κ, and Cohen's κ on binary preferences. Statistical significance was denoted as $^{*}P < 0.05$, $^{**}P < 0.01$, and $^{***}P < 0.001$.

Author Contributions

SAL, ZE, AAH, AAM contributed to the conception and design of the work; SAL, ZE, MZ, MG, SS, AS, PR, DM, AAH, AAM contributed to the data acquisition and curation; SAL, ZE, MZ, MG, PR, MZP, MAX, DM, AZF, AAH, AAM contributed to the technical implementation; HP, MZP, DM, DA, YY, JNC, AZF, LS, MM, SP, AAH, AAM provided technical and infrastructure guidance; SS, AZF, LS provided clinical inputs to the study; SAL, ZE, AAH contributed to the supplementary data analysis; SAL, ZE, MZ, MG, SS, AS, HP, PR, MZP, MAX, DM, DA, YY, JNC, AZF, LS, MM, SP, AAH, AAM contributed to the drafting and revising of the manuscript.

Data availability

All of Us data are accessible to registered researchers through the *All of Us* Researcher Workbench (https://www.researchallofus.org/). The WEAR-ME dataset is available to approved researchers for reproducibility purposes (access instructions at https://github.com/Google-Health/consumer-health-research/tree/main/insulin_resistance_prediction). De-identified longitudinal Fitbit records cannot be publicly shared due to participant consent and privacy restrictions.

Code availability

Code for Coachable Health Age, SensorWM, and all statistical analyses will be released upon publication at https://github.com/Google-Health/consumer-health-research.

Acknowledgements

We would like to thank our co-authors and collaborators from Google Research and academia for their contributions to this work. We would also like to thank the users who consented their data to make this research possible. We additionally would like to thank the clinicians and lifestyle coaches who provided their expertise to make this work possible.

Correspondence

Correspondence should be addressed to {aliheydari, aametwally}@google.com

References

[1] Donald M Lloyd-Jones, Norrina B Allen, Cheryl AM Anderson, Terrie Black, LaPrincess C Brewer, Randi E Foraker, Michael A Grandner, Helen Lavretsky, Amanda Marma Perak, Garima Sharma, et al. Life's essential 8: updating and enhancing the american heart association's construct of cardiovascular health: a presidential advisory from the american heart association. *Circulation*, 146(5):e18–e43, 2022.

[2] Muthiah Vaduganathan, George A. Mensah, Joseph V. Turco, Valentin Fuster, and Gregory A. Roth. The global burden of cardiovascular diseases and risks: A compass for future health. *Journal of the American College of Cardiology*, 80(25):2361–2371, 2022.

[3] Petr Klemera and Stanislav Ďoubal. A new approach to the concept and computation of biological age. *Mechanisms of Ageing and Development*, 127(3):240–248, 2006.

[4] Morgan E Levine, Ake T Lu, Austin Quach, Brian H Chen, Themistocles L Assimes, Stefania Bandinelli, Lifang Hou, Andrea A Baccarelli, James D Stewart, Yun Li, et al. An epigenetic biomarker of aging for lifespan and healthspan. *Aging (albany NY)*, 10(4):573, 2018.

[5] Andrew C. Miller, Joseph Futoma, Salar Abbaspourazad, Christina Heinze-Deml, Saba Emrani, Ian Shapiro, and Guillermo Sapiro. A wearable-based aging clock associates with disease and behavior. *Nature Communications*, 16(1):9264, oct 2025.

[6] David Spiegelhalter. How old are you, really? communicating chronic risk through 'effective age' of your body and organs. *BMC Medical Informatics and Decision Making*, 16:104, 2016.

[7] Carissa Bonner, Carys Batcup, Samuel Cornell, Michael Anthony Fajardo, Anna L Hawkes, Lyndal Trevena, and Jenny Doust. Interventions using heart age for cardiovascular disease risk communication: Systematic review of psychological, behavioral, and clinical effects. *JMIR Cardio*, 5(2):e31056, 2021.

[8] Gary Parkes, Trisha Greenhalgh, Mark Griffin, and Richard Dent. Effect on smoking quit rate of telling patients their lung age: the Step2quit randomised controlled trial. *BMJ*, 336(7644):598–600, mar 2008.

[9] Mina Bakhit, Samantha Fien, Eman Abukmail, Mark Jones, Justin Clark, Anna Mae Scott, Paul Glasziou, and Magnolia Cardona. Cardiovascular disease risk communication and prevention: a meta-analysis. *European Heart Journal*, 45(12):998–1011, 2024.

[10] Steve Horvath. Dna methylation age of human tissues and cell types. *Genome Biology*, 14(10):R115, 2013.

[11] Timothy V Pyrkov, Evgeny Getmantsev, Boris Zhurov, Konstantin Avchaciov, Mikhail Pyatnitskiy, Leonid Menshikov, Kristina Khodova, Andrei V Gudkov, and Peter O Fedichev. Quantitative characterization of biological age and frailty based on locomotor activity records. *Aging (Albany NY)*, 10(10):2973–2990, 2018.

[12] Cathal McCrory, Giovanni Fiorito, Belinda Hernandez, Silvia Polidoro, Aisling M O'Halloran, Ann Hever, Cliona Ni Cheallaigh, Ake T Lu, Steve Horvath, Paolo Vineis, et al. Grimage outperforms other epigenetic clocks in the prediction of age-related clinical phenotypes and all-cause mortality. *The Journals of Gerontology: Series A*, 76(5):741–749, 2021.

[13] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11):e745–e750, 2021.

[14] Susan Michie, Charles Abraham, Craig Whittington, John McAteer, and Sunjai Gupta. Effective techniques in healthy eating and physical activity interventions: a meta-regression. *Health Psychology*, 28(6):690–701, 2009.

[15] Susan Michie, Michelle Richardson, Marie Johnston, Charles Abraham, Jill Francis, Wendy Hardeman, Martin P Eccles, James Cane, and Caroline E Wood. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. *Annals of Behavioral Medicine*, 46(1):81–95, 2013.

[16] Albert Bandura. Health promotion by social cognitive means. *Health Education & Behavior*, 31(2):143–164, 2004.

[17] Leroy Hood and Mauricio Flores. A personal view on systems medicine and the emergence of proactive p4 medicine: predictive, preventive, personalized and participatory. *New Biotechnology*, 29(6):613–624, 2012.

[18] Alvin Rajkomar, Eyal Oren, Kai Chen, et al. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1(1):18, 2018.

[19] Salar Abbaspourazad, Oussama Elachqar, Andrew Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. Large-scale training of foundation models for wearable biosignals. In *The Twelfth International Conference on Learning Representations*.

[20] 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.

[21] Simon A. Lee, Cyrus Tanade, Hao Zhou, Juhyeon Lee, Megha Thukral, Md Sazzad Hissain Khan, Keum San Chun, Baiying Lu, Migyeong Gwak, Mehrab Bin Morshed, Viswam Nathan, Md Mahbubur Rahman, Li Zhu, Subramaniam Venkatraman, and Sharanya Arcot Desai. HiMAE: Hierarchical masked autoencoders discover resolution-specific structure in wearable time series. In *The Fourteenth International Conference on Learning Representations*, 2026.

[22] Hao Zhou, Simon A Lee, Cyrus Tanade, Keum San Chun, Juhyeon Lee, Migyeong Gwak, Megha Thukral, Justin Sung, Eugene Hwang, Mehrab Bin Morshed, et al. Physiology-aware masked cross-modal reconstruction for biosignal representation learning. *arXiv preprint arXiv:2605.00973*, 2026.

[23] Peng Cao, Zhijian Yang, Tennison Liu, Jonathan Wang, Jiang Wu, Magdalena Proszewska, Arvind Pillai, Mingwu Gao, Amir Farjadian, Lawrence Cai, et al. Wavesfm: Hierarchical representation learning for longitudinal wearable sensor waveforms. *arXiv preprint arXiv:2605.09173*, 2026.

[24] Zilin Jing, Vincent Jeanselme, Yuta Kobayashi, Simon A Lee, Chao Pang, Aparajita Kashyap, Yanwei Li, Xinzhuo Jiang, and Shalmali Joshi. One loss to rule them all: Marked time-to-event for structured EHR foundation models. In *The Fortieth Annual Conference on Neural Information Processing Systems*, 2026.

[25] Zechen Li, Keerthana Natarajan, Weizhi Zhang, Menglian Zhou, Simon A Lee, Yuwei Zhang, Maxwell A Xu, Zeinab Esmaeilpour, Flora D Salim, Mark Malhotra, et al. Glucofm: A dual-stream foundation model for continuous glucose monitoring. *arXiv preprint arXiv:2605.30865*, 2026.

[26] 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, pages 59382–59413, 2025.

[27] Evangelia Katsoulakis, Qi Wang, Huanmei Wu, et al. Digital twins for health: a scoping review. *npj Digital Medicine*, 7(1):77, 2024.

[28] Simon A Lee, Jiahang Hank Sha, Soomin Chung, Kai Akamatsu, Eran Halperin, and Jeffrey N Chiang. Forecasting patient trajectories from the electronic health records with action-conditioned world models. *arXiv preprint arXiv:2610.23862*, 2026.

[29] Bryan Lim, Ahmed M. Alaa, and Mihaela van der Schaar. Forecasting treatment responses over time using recurrent marginal structural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[30] Ioana Bica, Ahmed M. Alaa, James Jordon, and Mihaela van der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *International Conference on Learning Representations (ICLR)*, 2020.

[31] Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes. In *International Conference on Machine Learning (ICML)*, 2022. arXiv:2204.07258.

[32] Sarah M Urbut, Yi Ding, Tetsushi Nakao, Satoshi Koyama, Anika Misra, Xilin Jiang, Achyutha Harish, Leslie Gaffney, Whitney E Hornsby, Jordan W Smoller, et al. A bayesian framework for longitudinal ehr and genetic discovery. *Nature*, pages 1–12, 2026.

[33] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. Published as "Recurrent World Models Facilitate Policy Evolution," NeurIPS 2018.

[34] Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

[35] Mike A Merrill, Akshay Paruchuri, Naghmeh Rezaei, Geza Kovacs, Javier Perez, Yun Liu, Erik Schenck, Nova Hammerquist, Jake Sunshine, Shyam Tailor, et al. Transforming wearable data into personal health insights using large language model agents. *Nature Communications*, 17(1):1143, 2026.

[36] Justin Khasentino, Anastasiya Belyaeva, Xin Liu, Zhun Yang, Nicholas A Furlotte, Chace Lee, Erik Schenck, Yojan Patel, Jian Cui, Logan Douglas Schneider, et al. A personal health large language model for sleep and fitness coaching. *Nature Medicine*, 31(10):3394–3403, 2025.

[37] A Ali Heydari, Ken Gu, Vidya Srinivas, Hong Yu, Zhihan Zhang, Yuwei Zhang, Akshay Paruchuri, Qian He, Hamid Palangi, Nova Hammerquist, et al. The anatomy of a personal health agent. *arXiv preprint arXiv:2508.20148*, 2025.

[38] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature Medicine*, 29(8):1930–1940, 2023.

[39] Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A Fries, Michael Wornow, Akash Swaminathan, L Soleymani Lehmann, et al. Testing and evaluation of health care applications of large language models: A systematic review. *JAMA*, 333(4):319–328, 2025.

[40] Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. Leworld-model: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026.

[41] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting, 2024.

[42] Mahdi Moqri, Chiara Herzog, Jesse R Poganik, Jamie Justice, Daniel W Belsky, Albert Higgins-Chen, Alexey Moskalev, Georg Fuellen, Alan A Cohen, Ivan Bautmans, et al. Biomarkers of aging for the identification and evaluation of longevity interventions. *Cell*, 186(18):3758–3775, 2023.

[43] Gregory Hannum, Justin Guinney, Ling Zhao, LI Zhang, Guy Hughes, SriniVas Sadda, Brandy Klotzle, Marina Bibikova, Jian-Bing Fan, Yuan Gao, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular cell*, 49(2):359–367, 2013.

[44] Ake T Lu, Austin Quach, James G Wilson, Alex P Reiner, Abraham Aviv, Kenneth Raj, Lifang Hou, Andrea A Baccarelli, Yun Li, James D Stewart, et al. Dna methylation grimage strongly predicts lifespan and healthspan. *Aging (albany NY)*, 11(2):303, 2019.

[45] Daniel W Belsky, Avshalom Caspi, David L Corcoran, Karen Sugden, Richie Poulton, Louise Arseneault, Andrea Baccarelli, Kartik Chamarti, Xu Gao, Eilis Hannon, et al. Dunedinpace, a dna methylation biomarker of the pace of aging. *elife*, 11:e73420, 2022.

[46] Segterra, Inc. *InnerAge 2.0 Blood Biomarker Aging Model*. InsideTracker, 2021. U.S. Patent 8762167 and patent pending applications.

[47] 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, 2008.

[48] Garmin Ltd. *Garmin Fitness Age and Firstbeat Analytics VO2 max Population Norms*. Garmin International, Inc., 2021. Based on Cooper Institute VO2 max population percentiles.

[49] Ōura Health. Beyond heart rate: Cardiovascular age unlocks new dimensions of health. Technical white paper, Ōura Health Oy, May 2024. Focuses on PPG-derived Pulse Wave Velocity (PWV) and arterial stiffness.

[50] Andreas Brink-Kjaer, Eileen B Leary, Haoqi Sun, M Brandon Westover, Katie L Stone, Paul E Peppard, Nancy E Lane, Peggy M Cawthon, Susan Redline, Poul Jennum, et al. Age estimation from sleep studies using deep learning predicts life expectancy. *NPJ digital medicine*, 5(1):103, 2022.

[51] WHOOP Inc. and Buck Institute for Research on Aging. Whoop healthspan: Whoop age and pace of aging (all-cause mortality hazard ratio modeling). Scientific white paper, WHOOP Inc. & Buck Institute for Research on Aging, September 2025. Foreword by Eric Verdin, CEO of the Buck Institute.

[52] Yanjun Li, Qi Huang, Jin Jiang, Xusheng Du, Wenxin Xiang, Shiqi Zhang, Zean Pan, Liyuan Zhao, Yuyan Cui, Limei Ke, et al. Large language model-based biological age prediction in large-scale populations. *Nature Medicine*, 31(9):2977–2990, 2025.

[53] Indraneel Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with shapley-value-based explanations as feature importance measures. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119, pages 5491–5500. PMLR, 2020.

[54] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 353–362. ACM, 2021.

[55] Dongfeng Zhang, Xiaoli Shen, and Xin Qi. Resting heart rate and all-cause and cardiovascular mortality in the general population: a meta-analysis. *CMAJ*, 188(3):E53–E63, Feb 2016.

[56] H. Master, J. Annis, S. Huang, J. A. Beckman, F. Ratsimbazafy, K. Marginean, R. Carroll, K. Natarajan, F. E. Harrell, D. M. Roden, P. Harris, and E. L. Brittain. Association of step counts over time with the risk of chronic disease in the all of us research program. *Nature Medicine*, 28(11):2301–2308, 2022.

[57] Iyas Daghlas, Hassan S. Dashti, Jacqueline Lane, Krishna G. Aragam, Martin K. Rutter, Richa Saxena, and Céline Vetter. Sleep duration and myocardial infarction. *Journal of the American College of Cardiology*, 74(10):1304–1314, 2019.

[58] S. Kany, M. A. Al-Alusi, J. T. Rämö, J. P. Pirruccello, T. W. Churchill, S. A. Lubitz, M. Maddah, J. S. Guseh, P. T. Ellinor, and S. Khurshid. Associations of "weekend warrior" physical activity with incident disease and cardiometabolic health. *Circulation*, 150(16):1236–1247, 2024.

[59] Jiehua Wei, Minxue Shen, Shenxin Li, Yi Xiao, Dan Luo, Gerson Ferrari, Dong Hoon Lee, Leandro FM Rezende, Jason MR Gill, Matthew N Ahmadi, et al. Volume vs intensity of physical activity and risk of cardiovascular and non-cardiovascular chronic diseases. *European Heart Journal*, page ehag168, 2026.

[60] 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 Network Open*, 1(6):e183605, 2018.

[61] Tianyi Huang, Sara Mariani, and Susan Redline. Sleep irregularity and risk of cardiovascular events: The multi-ethnic study of atherosclerosis. *Journal of the American College of Cardiology*, 75(9):991–999, 2020.

[62] Ahmed A Metwally, A Ali Heydari, Daniel McDuff, Alexandru Solot, Zeinab Esmaeilpour, Anthony Z Faranesh, Menglian Zhou, Girish Narayanswamy, Maxwell A Xu, Xin Liu, et al. Insulin resistance prediction from wearables and routine blood biomarkers. *Nature*, 652(8109):451–461, 2026.

[63] Matthew J. Gurka, Christa L. Lilly, Mark N. Oliver, and Mark D. DeBoer. An examination of sex and racial/ethnic differences in the metabolic syndrome among adults: a confirmatory factor analysis and a resulting continuous severity score. *Metabolism*, 63(2):218–225, 2014.

[64] Stephen M Smith, Diego Vidaurre, Fidel Alfaro-Almagro, Thomas E Nichols, and Karla L Miller. Estimation of brain age delta from brain imaging. *NeuroImage*, 200:528–539, 2019.

[65] Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, et al. Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys*, 58(3):1–38, 2025.

[66] Matthew McDermott, Bret Nestor, Peniel Argaw, and Isaac S Kohane. Event stream gpt: a data pre-processing and modeling library for generative, pre-trained transformers over continuous-time sequences of complex events. *Advances in neural information processing systems*, 36:24322–24334, 2023.

[67] Simon A Lee, Sujay Jain, Alex Chen, Kyoka Ono, Arabdha Biswas, Ákos Rudas, Jennifer Fang, and Jeffrey N Chiang. Clinical decision support using pseudo-notes from multiple streams of ehr data. *npj Digital Medicine*, 8(1):394, July 2025.

[68] Matthew BA McDermott, Ethan Steinberg, Jason A Fries, Robin P van de Water, Chao Pang, Patrick Rockenschaub, Pawel Renc, Jungwoo Oh, Kamilė Stankevičiūtė, Justin Xu, et al. Meds—an emerging data standard and ecosystem for health ai research. *NEJM AI*, 3(6):AIra2501253, 2026.

[69] Greg Atkinson and Damien Davenne. Relationships between sleep, physical activity and human health. *Physiology & behavior*, 90(2-3):229–235, 2007.

[70] Daniel J Buysse, Ron Grunstein, Jim Horne, and Peretz Lavie. Can an improvement in sleep positively impact on health? *Sleep medicine reviews*, 14(6):405–410, 2010.

[71] Kenneth E Powell, Amanda E Paluch, and Steven N Blair. Physical activity for health: What kind? how much? how intense? on top of what? *Annual review of public health*, 32(1):349–365, 2011.

[72] Ying Wang, Oumayma Bounou, Gaoyue Zhou, Randall Balestriero, Tim GJ Rudner, Yann LeCun, and Mengye Ren. Temporal straightening for latent planning. *arXiv preprint arXiv:2603.12231*, 2026.

[73] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014.

[74] Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845, 1988.

[75] Donald L Kirkpatrick et al. Techniques for evaluating training programs. *Journal of the American Society of training Directors*, 13(11):3–9, 1959.

[76] Weizhi Zhang, Zechen Li, Hamid Palangi, Ben Graef, A Ali Heydari, Simon A Lee, Salman Rahman, Ray Luo, Zeinab Esmaeilpour, Erik Schenck, et al. Rubricstree: Scalable and evolving open-ended evaluation of personal health agents across health memory and medical skills. *arXiv preprint arXiv:2606.18203*, 2026.

[77] Joshua F Wiley and Melinda J Carrington. A metabolic syndrome severity score: A tool to quantify cardio-metabolic risk factors. *Preventive medicine*, 88:189–195, 2016.

[78] Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem, 2021.

[79] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[80] Rahul Dey and Fathi M Salem. Gate-variants of gated recurrent unit (gru) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, pages 1597–1600. IEEE, 2017.

Extended Data

Prospective cardiometabolic associations of Coachable Health Age

Extended Data Table 1 | **Kaplan–Meier cardiometabolic disease-free survival probabilities by baseline Coachable Health Age status.** Disease-free survival probabilities (95% CI) at 4, 8, 16, 24, and 32 quarters (1, 2, 4, 6, and 8 years) of follow-up comparing Slower/Normal Agers ($\Delta\text{Age} \leq 0$) and Accelerated Agers ($\Delta\text{Age} > 0$).

Follow-up Horizon	Slower/Normal Agers ($\Delta\text{Age} \leq 0$) Disease-Free Survival (95% CI)	Accelerated Agers ($\Delta\text{Age} > 0$) Disease-Free Survival (95% CI)	Absolute Risk Difference (Δ Survival, %)
4 Quarters (1 Year)	98.31% (98.03–98.54%)	97.70% (97.31–98.03%)	+0.61%
8 Quarters (2 Years)	96.72% (96.34–97.06%)	95.51% (94.95–96.00%)	+1.21%
16 Quarters (4 Years)	91.91% (91.24–92.53%)	90.35% (89.41–91.22%)	+1.55%
24 Quarters (6 Years)	86.99% (86.00–87.91%)	83.72% (82.18–85.15%)	+3.26%
32 Quarters (8 Years)	82.41% (80.70–83.99%)	79.28% (77.03–81.34%)	+3.13%

Log-rank test comparing Slower/Normal Agers ($n = 10,697$) vs. Accelerated Agers ($n = 7,726$): $\chi^2 = 20.17$, $p = 7.09 \times 10^{-6}$.

Extended Data Table 2 | **Multivariable Cox proportional hazards regression models for incident cardiometabolic disease.** Log-hazard ratios ($\hat{\beta} \pm$ SE), hazard ratios (HR) with 95% confidence intervals (CI), z-statistics, and two-sided p-values for (A) the lagged time-varying Cox model ($N = 13,842$ participants; 164,336 three-month intervals; 906 events) where ΔAge_{t-1} predicts incident diagnosis in window t, and (B) the baseline landmark Cox model ($N = 14,663$ participants; 1,402 events). Both models are adjusted for chronological age at baseline, baseline BMI, and sex.

Model & Covariate	Log-HR ($\hat{\beta} \pm$ SE)	HR (95% CI)	z-stat	p-value
A. Lagged Time-Varying Cox Model (*N = 13,842**; 164,336 **intervals;** 906 **events)***[a]*				
Lagged Health Age Delta (ΔAge_{t-1}, +1 yr)	0.04 ± 0.01	**1.04** (1.02–1.05)	4.99	6.0×10^{-7}
Baseline chronological age (+1 yr)	0.02 ± 0.002	**1.03** (1.02–1.03)	10.56	$< 10^{-16}$
Baseline BMI (+1 kg/m^2)	0.03 ± 0.003	**1.03** (1.03–1.04)	11.18	$< 10^{-16}$
Sex: Male (vs. Female)	0.09 ± 0.07	1.09 (0.94–1.27)	1.20	0.23
Sex: Other / Prefer not to answer / Skipped	−0.04 ± 0.25	0.96 (0.59–1.58)	−0.15	0.88
B. Baseline Landmark Cox Model (*N = 14,663**; 1,402 **events)***[b]*				
Baseline Health Age Delta (ΔAge_0, +1 yr)	0.02 ± 0.01	**1.02** (1.01–1.04)	3.63	2.8×10^{-4}
Baseline chronological age (+1 yr)	0.02 ± 0.002	**1.02** (1.02–1.03)	11.72	$< 10^{-16}$
Baseline BMI (+1 kg/m^2)	0.05 ± 0.004	**1.05** (1.04–1.05)	12.20	$< 10^{-16}$
Sex: Male (vs. Female)	0.02 ± 0.06	1.02 (0.91–1.15)	0.37	0.71
Sex: Other / Prefer not to answer / Skipped	−0.27 ± 0.22	0.76 (0.49–1.17)	−1.24	0.21

[a]Partial AIC = 15,657.39; log-likelihood ratio test = 231.45 on 5 df ($p < 10^{-16}$). [b]Partial AIC = 24,648.38; log-likelihood ratio test = 317.80 on 5 df ($p < 10^{-16}$).

Extended Data Table 3 | **Incremental prognostic discrimination of Coachable Health Age beyond routine clinical demographics.** Comparison of nested Cox proportional hazards models predicting incident cardiometabolic disease, with and without Coachable Health Age, in the disease-free baseline cohort. Discrimination was quantified using Harrell's C-index, with 95% confidence intervals. Differences in C-index (ΔC) were evaluated using paired bootstrap resampling.

Cohort (N; Events)	**Baseline Clinical (Age + Sex + BMI)**	**+ Coachable Health Age**	**Δ C-index [95% CI]**	**p-value**
Full Cohort (14,663; 1,402)	0.6485 [0.6340–0.6632]	**0.6541** [0.6390–0.6685]	**+0.0056** [+0.0018, +0.0096]	0.005
5-Fold CV (14,663; 1,402)	0.6475 [0.6330–0.6615]	**0.6529** [0.6390–0.6668]	**+0.0053** [+0.0012, +0.0098]	0.016
Adults < 50 yr (7,399; 549)	0.6648 [0.6420–0.6910]	**0.6762** [0.6525–0.7010]	**+0.0114** [+0.0048, +0.0181]	< 0.001

Extended Data Table 4 | **Kirkpatrick-Inspired Evaluation Rubric.** The automated evaluation framework comprises four levels and twelve dimensions, each assessed using a five-point anchored rating scale.

Evaluation Level	**Guiding Question**	**Assessment Dimensions**
L1. Reaction	*Will the user engage?*	**1a.** Clarity and readability **1b.** Tone and motivation **1c.** Personal relevance
L2. Learning	*Does the user understand?*	**2a.** Data comprehension **2b.** Causal understanding **2c.** Insight and stakes ("why now")
L3. Behavior	*Will the user follow it?*	**3a.** Actionability and specificity **3b.** Feasibility and progression **3c.** Behavior-change technique
L4. Results	*Is improvement likely?*	**4a.** Predicted health impact **4b.** Measurability **4c.** Evidence grounding

Extended Data Table 5 | **Comparison of counterfactual and generic recommendations across evaluation dimensions.** Mean scores for each of the 12 rubric dimensions, evaluated on a five-point scale (1 = lowest, 5 = highest) across 2,613 paired observations. Of 2,780 eligible participants with at least 3 weeks of wearable data, 155 were excluded prior to simulation due to missing (zero-imputed) baseline sleep duration in the final observation week and 12 were excluded due to unparseable judge outputs, yielding 2,613 paired evaluations. Scores are reported before application of the safety gate. Boldface highlights dimensions exhibiting particularly large differences between the counterfactual and generic approaches. CF, Counterfactual Coaching.

Dim	Rubric dimension	CF	Generic
1a	Clarity & readability	5.00	5.00
1b	Tone & motivation	5.00	4.96
1c	Personal relevance	4.99	4.40
2a	Data comprehension	4.99	4.22
2b	Causal understanding	4.98	3.76
2c	**Insight / stakes ("why now")**	**4.98**	**2.86**
3a	Actionability & specificity	4.92	4.65
3b	Feasibility & progression	4.95	4.27
3c	Behavior-change technique	4.98	4.75
4a	Predicted health impact	4.94	4.17
4b	Measurability	4.89	4.62
4c	**Evidence grounding**	**4.92**	**4.15**

Extended Data Table 6 | **Within-participant comparison of counterfactual and generic recommendations.** The pre-specified contrast is evaluated using weighted overall scores from paired observations ($n = 2,613$). Δ denotes the mean paired difference, with positive values indicating higher scores for counterfactual recommendations; d_T is Cohen's effect size for paired observations. Statistical significance is assessed using a Wilcoxon signed-rank test. Win, tie and loss counts indicate the number of paired comparisons in which counterfactual recommendations scored higher, equally or lower, respectively. Preference denotes the proportion of pooled votes favoring counterfactual recommendations in blinded head-to-head evaluations.

Contrast	Δ	d_T	p	W / T / L	Preferred
Counterfactual → Generic	+0.70	1.86	$< 10^{-300}$	2,570 / 3 / 40	98.2%

Appendix Table of Contents

A Supplementary Information **50**

A.1 Coachable Health Age 50

A.2 Trajectory Prediction with Sensor World Model 51

A.3 Expert Evaluation of Counterfactual Coaching 54

B Supplementary Figures **58**

B.1 Coachable Health Age 58

B.2 Trajectory Prediction with Sensor World Model 60

C Supplementary Tables **65**

C.1 Coachable Health Age 65

C.2 Trajectory Prediction with Sensor World Model 67

C.3 Expert Evaluation of Counterfactual Coaching 77

A. Supplementary Information

A.1. Coachable Health Age

LLM Health Age & Consistency

Within-Temperature Consistency The stochastic stability of the LLM's estimations was evaluated by comparing the consistency of health age predictions across three runs ($k = 3$) under different temperatures ($T \in \{0.1, 0.5, 1.0\}$) for two data input configurations: (1) Raw Signals (incorporating all raw wearable sensor streams) and (2) Identical metrics that coachable Health Age algorithm receives as input. In order to be able to compare across temperatures, in cases where LLM provided a range as health age (e.g., 45-48 years), the average of the interval was used in the analysis. For the Raw Signals configuration, the model exhibited greater stability across temperatures (Supplementary Table 3). At $T = 0.1$, the average standard deviation (SD) between runs was low (1.35 years). Single-run consistency, measured by ICC(3,1), was moderate (0.54), and averaging across the three runs improved consistency to ICC$(3, 3) = 0.78$. As temperature increased, consistency declined gradually: at $T = 0.5$, average SD rose to 2.06 years (ICC$(3, 1) = 0.30$, ICC$(3, 3) = 0.57$), and at $T = 1.0$, SD reached 3.37 years while retaining higher reliability than the curated baseline (ICC$(3, 1) = 0.20$, ICC$(3, 3) = 0.42$).

The Identical CHA configuration showed a sharp drop in consistency as temperature increased (Supplementary Table 3). While reliability at $T = 0.1$ matched the all-signals configuration in ICC (ICC$(3, 1) = 0.54$, ICC$(3, k) = 0.78$) albeit with higher run-to-run spread (SD = 2.72 years), reliability collapsed at $T = 0.5$ (SD = 3.54 years), with single-run consistency dropping to ICC$(3, 1) = 0.09$ and three-run average consistency only reaching ICC$(3, k) = 0.23$. A similar collapse was observed at $T = 1.0$ (SD = 3.20 years, ICC$(3, 1) = 0.11$, ICC$(3, k) = 0.28$).

Across-Temperature Sensitivity For the **Raw Signals** configuration, the Friedman test indicated a statistically significant overall difference in health age estimations across the three temperatures ($\chi^2(2) = 6.47, p = 0.039$). However, post-hoc Wilcoxon signed-rank tests with Bonferroni correction did not detect any statistically significant differences between specific temperature pairs ($T_{0.1}$ vs. $T_{0.5}$: adj. $p = 0.075$; $T_{0.5}$ vs. $T_{1.0}$: adj. $p = 0.0689$; $T_{0.1}$ vs. $T_{1.0}$: adj. $p = 1.00$).

For the **Curated Metrics** configuration, the sensitivity to temperature was highly pronounced. The Friedman test revealed an extremely significant difference across temperatures ($\chi^2(2) = 39.71, p = 2.39 \times 10^{-9}$). Post-hoc Wilcoxon signed-rank tests with Bonferroni correction confirmed that estimated health ages systematically shifted across all temperature pairs ($T_{0.1}$ vs. $T_{0.5}$: adj. $p = 0.0039$; $T_{0.5}$ vs. $T_{1.0}$: adj. $p = 0.0024$; $T_{0.1}$ vs. $T_{1.0}$: adj. $p = 7.71 \times 10^{-10}$; Supplementary Table 4).

Association with standard Metrics (Modified LE8 and MSSS): To evaluate the clinical alignment of the generated health ages, we calculated the Pearson correlation coefficient (r) between the delta health age (estimated health age minus actual age) and two established health indicators: the modified Life's Essential 8 (mod-LE8) score and the Metabolic Syndrome Severity (MSSS) score. Across both data configurations, delta health age correlated negatively with the modified LE8 score and positively with the MSSS score (Supplementary Table 5).

Coachable Health Age Correlates with AHA Life's Essentials 8 & Metabolic Syndrome Score

We further evaluated the clinical alignment of Coachable Health Age by examining its associations with established measures of cardiovascular and metabolic health. The American Heart Association's Life's Essential 8 (LE8) framework [1] provides a comprehensive assessment of cardiovascular health across eight domains, including diet, physical activity, nicotine exposure, sleep health, body mass index, blood lipids, blood glucose, and blood pressure. We considered a modified LE8 score (mod-LE8) excluding diet, alongside the Metabolic Syndrome Severity Score (MSSS) [77], a continuous measure of metabolic dysfunction incorporating BMI, blood pressure, triglycerides, HDL cholesterol, and fasting glucose. CHA demonstrated the strongest absolute correlations with both mod-LE8 ($|r| = 0.512$) and MSSS ($|r| = 0.355$), outperforming chronological age and alternative biological age estimates. These findings suggest that CHA captures clinically meaningful variation in cardiometabolic health beyond that reflected by conventional age measures.

We additionally examined the relationship between wearable-derived Delta Health Age and self-reported overall health status. Consistent with the expected relationship between perceived health and physiological

aging, Delta Health Age increased monotonically with worsening self-reported health, from approximately zero years among individuals reporting excellent health to over 11 years among those reporting poor health. This graded association provides complementary evidence of construct validity, suggesting that wearable-derived health age estimates reflect meaningful differences in overall health status that are consistent with individuals' subjective assessments of their health.

A.2. Trajectory Prediction with Sensor World Model

Choosing the SensorWM

We compared two world model training formulations, Trajectory Transformer [78], which autoregressively predicts future wearable representations using a transformer backbone [79], and LeWorldModel [40], which learns action-conditioned latent dynamics through joint-embedding prediction. To determine which formulation better preserved clinically relevant information, we evaluated downstream classification performance using generated trajectories from both models against observed Fitbit data. LeWorldModel outperformed Trajectory Transformer on nine of 12 tasks, with notable improvements in insulin resistance (AUROC 0.760 vs. 0.719), smoking status (0.774 vs. 0.698), and hyperlipidemia (0.692 vs. 0.645). Based on its stronger preservation of clinically informative representations, we adopted the LeWorldModel-inspired training formulation for SensorWM (Supplementary Table 13).

Scaling SensorWM

We investigated the effect of pre-training data scale on SensorWM's ability to generate clinically informative physiological trajectories, comparing models trained on 1k, 10k, and 100k participants. We evaluated representation quality indirectly through downstream classification performance using model-generated trajectories. As shown in Supplementary Table 14, increasing pre-training data generally improved downstream discrimination, with the 100k model achieving the highest AUROC on 8 of 12 tasks, including insulin resistance (0.760), respiratory conditions (0.660), and smoking status (0.774). Although improvements were not strictly monotonic across all tasks, these results suggest that scaling pre-training data enhances the preservation of clinically relevant information in generated physiological trajectories, motivating our use of the 100k-participant model for SensorWM.

Trajectory versus Static Classification

While world models are trained to forecast future physiological states rather than optimize directly for discriminative downstream classification, an important question is whether the longitudinal dynamics captured by the model encode information that is absent from static snapshots of wearable data. To investigate this, we compare downstream prediction performance using four different feature sources: (1) the true observed longitudinal Fitbit trajectories, (2) trajectories generated by the sensor world model, (3) representations obtained from the state of the art SensorFM [20], and (4) features derived solely from the first week of wearable observations.

This comparison isolates the value of longitudinal physiological information. Both the SensorFM and first-week baselines operate on a limited temporal view of an individual's health state and therefore primarily capture cross-sectional physiological characteristics at a single point in time. In contrast, the observed Fitbit trajectories and the world-model-generated trajectories incorporate patterns that emerge over extended periods, including behavioral consistency, physiological adaptation, recovery dynamics, and temporal variability. These longitudinal signatures may encode information about an individual's health that cannot be recovered from a single observation window alone helping motivate the need for having a world model capable of learning the physiological trajectories of human health. All models utilize the same Gated Recurrent Unit (GRU) style classifier [80] to ensure fairness and Longitudinal Models (True Fitbit Sequence & Sensor World Model Fitbit Sequences) and SensorFM [20] latent space dimensions are fixed at $D = 128$ to ensure fair comparisons.

We therefore evaluate whether access to a longitudinal trajectory provides measurable benefits for downstream health prediction. If longitudinal physiological dynamics contain information beyond static physiological snapshots, models trained on observed or generated trajectories should outperform approaches that rely solely on single-time-point measurements. Furthermore, comparable performance between observed and generated trajectories would suggest that the sensor world model successfully preserves the clinically

relevant information contained within real wearable trajectories while enabling future-state simulation and forecasting. This analysis therefore serves both as an assessment of the utility of longitudinal wearable signals and as a validation of the information fidelity of the learned world model representations.

Classification Subgroup Analysis

To evaluate the fidelity of the generated physiological trajectories, we perform a series of subgroup analyses constructed from our demographic labels stratifying by age, body mass index (BMI), and sex. The objective of these experiments is not to assess demographic fairness, but rather to determine whether SensorWM preserves clinically relevant physiological information across populations that exhibit substantially different baseline physiology, behavioral patterns, and disease prevalence. If the generated trajectories faithfully capture the underlying structure of the observed wearable signals, downstream predictive performance should remain similar when using world-model-generated features instead of the true wearable measurements.

For each demographic subgroup, we repeat the downstream prediction experiments using both observed Fitbit features and world-model-generated features. Performance is quantified using AUROC, and differences between the two feature sources are evaluated using DeLong's test. By comparing prediction performance within individual demographic strata, we can directly assess whether the forecasting process preserves the predictive signal contained in the original wearable measurements. Large performance degradations would suggest that information relevant to a particular population is lost during forecasting, whereas comparable performance indicates that the generated trajectories maintain the physiological characteristics necessary for downstream inference.

This stratified evaluation provides a particularly stringent assessment of feature fidelity because the world model must accurately reproduce physiological dynamics across populations with distinct health profiles and behavioral patterns. Consistent agreement between models trained on observed and generated features would indicate that the learned latent dynamics capture physiologically meaningful structure and that forecasted trajectories retain much of the information present in the original wearable observations. Such preservation is essential if generated physiological states are to serve as effective surrogates for future wearable measurements in downstream health applications.

Sensitivity Analysis of Error Propagation

A fundamental challenge in world model development is understanding how forecasting errors propagate into downstream clinical predictions. Although trajectory forecasting models are typically optimized using reconstruction objectives, improvements in reconstruction accuracy do not necessarily translate into improvements in clinically relevant decision-making. We therefore performed a sensitivity analysis to quantify the relationship between trajectory forecasting error and downstream prediction consistency across all evaluation tasks.

For each participant, we first generated a future wearable trajectory using the sensor world model and computed the trajectory reconstruction error relative to the observed wearable measurements. We then applied identical downstream classifiers to both the observed and world-model-generated trajectories and calculated the absolute difference in predicted probabilities between the two predictions. This analysis provides a direct measure of error propagation, allowing us to assess whether inaccuracies in trajectory forecasting lead to meaningful changes in downstream clinical inference. Participants were further categorized according to whether the classifier predictions obtained from observed and generated trajectories were concordant or discordant.

Across nearly all prediction tasks, we observed that downstream prediction differences remained small even when trajectory reconstruction errors varied substantially (Supplementary Figure 5). Most participants clustered near the origin of the plots, indicating that generated trajectories frequently produced predictions that closely matched those obtained from the true wearable measurements. Notably, many individuals with moderate forecasting errors still exhibited highly consistent downstream predictions, suggesting that reconstruction accuracy alone is not the primary determinant of downstream utility. Instead, the world model appears to preserve the physiologically relevant structure necessary for clinical prediction despite imperfect reconstruction of individual measurements.

Discordant predictions were relatively uncommon and were concentrated among participants exhibiting larger differences in predicted probabilities. Importantly, these identified discordance pairs were not

always associated with the largest trajectory reconstruction errors. In several tasks, participants with comparatively low reconstruction error nonetheless crossed classification decision boundaries, whereas others with substantially larger forecasting errors remained correctly classified. This observation suggests that downstream sensitivity depends not only on the magnitude of forecasting error but also on where that error occurs within the physiological state space. Errors affecting features strongly associated with a particular outcome may have a disproportionately large effect on prediction, whereas reconstruction errors in less informative regions may have minimal clinical consequence.

These findings indicate that the relationship between forecasting accuracy and downstream performance is highly nonlinear. While trajectory reconstruction error provides a useful measure of forecasting quality, it does not fully characterize the utility of generated physiological trajectories for clinical prediction. From a health modeling perspective, preserving the latent physiological signals relevant to downstream outcomes may be more important than perfectly reconstructing every observed wearable measurement. The overall concentration of points near low probability differences across tasks therefore provides evidence that the sensor world model retains much of the clinically meaningful information present in real wearable trajectories, even when forecasting errors accumulate over extended prediction horizons.

Feature Fidelity: Correlations

A central question is whether generated trajectories preserve the physiological relationships present in real-world sensor data. While trajectory reconstruction metrics evaluate the accuracy of individual feature forecasts, they do not assess whether the model maintains the higher-order dependencies that govern interactions between wearable features. Preserving these dependencies is essential if generated trajectories are to serve as realistic surrogates for future wearable observations. We therefore evaluated feature fidelity by comparing the correlation structure of observed and world-model-generated wearable trajectories.

For each cohort, we computed pairwise correlation matrices across all physiological and behavioral features using both the observed Fitbit trajectories and the corresponding trajectories generated by the Sensor World Model. Correlations were calculated at the population level as well as within behaviorally distinct subgroups consisting of sedentary and highly active participants. To quantify deviations from the true physiological structure, we additionally computed difference matrices representing the change in pairwise correlations between observed and generated trajectories. This analysis evaluates whether the world model preserves the statistical organization of the wearable feature space rather than simply reproducing individual measurements.

Across all cohorts, the generated trajectories closely reproduced the correlation structure observed in the original wearable data (Supplementary Figure 6, 7). Many of the strongest physiological relationships, including interactions between cardiovascular measurements, sleep architecture variables, activity metrics, and stress-related features, were preserved in the generated trajectories. The overall visual similarity between observed and generated correlation matrices suggests that the world model captures underlying physiological dependencies that extend beyond individual feature reconstruction which is a well known property of world models [34].

The difference matrices further indicate that most pairwise relationships remain stable following trajectory generation. Although localized deviations are present, particularly among a subset of activity- and sleep-related measurements, these changes are small relative to the overall correlation structure. Importantly, preservation of feature relationships was observed not only within the full population but also within sedentary and highly active cohorts, which exhibit substantially different behavioral and physiological profiles. This suggests that the world model learns physiological dynamics that generalize across heterogeneous patterns of activity rather than relying on population-average correlations alone.

These findings provide complementary evidence to the downstream prediction experiments. Whereas classification analyses evaluate whether generated trajectories preserve task-relevant predictive information, correlation analyses assess whether they preserve the broader physiological organization of the wearable feature space. The strong agreement between observed and generated correlation structures suggests that the world model retains much of the higher-order information encoded within longitudinal wearable measurements. Consequently, the generated trajectories appear to preserve both local feature behavior and global physiological structure, supporting their use as realistic simulations of future health trajectories.

A.3. Expert Evaluation of Counterfactual Coaching

The main-text expert evaluation compared Counterfactual Coaching with Generic Coaching in a blinded, randomized, paired study of 49 WEAR-ME participant cases reviewed by five certified lifestyle coaches and five board-certified clinicians. Across both expert panels, the study comprised 490 paired plan evaluations. Detailed per-rater, rubric-level, cross-disciplinary, and persona-stratified results are reported below and in Supplementary Tables 25–30.

Head-to-head preference and anticipated adherence

Lifestyle coaches completed head-to-head side-by-side comparisons of Counterfactual and Generic Coaching plans for each participant. In the set displayed in Fig. 1e, 245 valid paired comparisons were available. Counterfactual Coaching was selected as the Best Overall Plan in 71.4% of evaluations (175/245; 95% CI, 65.5–76.7%), compared with 28.6% for Generic Coaching (70/245; 95% CI, 23.3–34.5%). This corresponded to a +42.9 percentage-point difference and a 2.50-fold preference ratio (two-sided exact binomial $P = 1.49 \times 10^{-11}$).

Lifestyle coaches also selected Counterfactual Coaching as having the Highest Likelihood of Long-Term Adherence in 71.0% of displayed evaluations (174/245; 95% CI, 65.0–76.3%), compared with 29.0% for Generic Coaching (71/245; $P = 3.72 \times 10^{-11}$).

Preference was directionally consistent across all five lifestyle coaches. Individual-rater Counterfactual win rates ranged from 53.1% to 98.0% for Best Overall Plan and from 44.9% to 98.0% for anticipated long-term adherence (Supplementary Table 30). When evaluations were aggregated by participant using majority consensus of at least three of five coaches, Counterfactual Coaching was preferred in 43 of 49 cases (87.8%; $P = 5.73 \times 10^{-8}$) for Best Overall Plan and in 44 of 49 cases (89.8%; $P = 7.60 \times 10^{-9}$) for Highest Likelihood of Long-Term Adherence.

Clinical safety and physiological accuracy

Clinical safety was independently assessed by five board-certified clinicians across 245 paired evaluations. Counterfactual plans passed the Safe Intensity Limits criterion in 99.2% of evaluations (243/245; 95% CI, 97.1–99.8%), compared with 97.6% of Generic plans (239/245; 95% CI, 94.8–98.9%). The difference was not statistically significant (Δ = +1.6 percentage points; exact McNemar $P = 0.289$; Supplementary Table 27).

Physiological claims were judged correct in 97.1% of Counterfactual plans (238/245) and 97.6% of Generic plans (239/245; $P = 1.00$). Clinicians also rated Trajectory Control highly in both conditions (96.7% [237/245] for Counterfactual and 93.1% [228/245] for Generic; Δ = +3.7 percentage points; $P = 0.093$). These results indicate that the additional projected physiological information in the Counterfactual condition did not produce a detectable reduction in safety or physiological correctness.

Lifestyle coach rubric performance

We analyzed paired top-box achievement across 10 lifestyle-coach evaluation dimensions (Supplementary Table 26). Counterfactual Coaching achieved statistically significant improvements across seven of the ten dimensions.

The largest difference occurred for Contrasts Benefits vs. Status Quo. Coaches rated 73.5% of Counterfactual plans (180/245) as effectively contrasting the projected benefits of behavioral change with maintaining current behavior, compared with 5.7% of Generic plans (14/245). This corresponded to a +67.8 percentage-point difference, with 170 discordant pairs favoring Counterfactual and 4 favoring Generic (exact McNemar $P = 3.15 \times 10^{-45}$).

Counterfactual Coaching also increased the proportion of plans judged to provide a Compelling Reason to Change without fear- or shame-based framing from 28.6% (70/245) to 75.5% (185/245), a +46.9 percentage-point difference (121 versus 6 discordant pairs; $P = 6.39 \times 10^{-29}$). Trajectory Control similarly increased from 59.4% (145/244) for Generic Coaching to 97.1% (238/245) for Counterfactual Coaching, with 97 versus 5 discordant pairs ($P = 3.46 \times 10^{-23}$).

On the ordinal behavior-change rubrics, Counterfactual Coaching more than doubled the frequency of

Strongly Applied Behavior-Change Strategies, from 20.4% (50/245) to 50.2% (123/245; Δ = +29.8 percentage points; 89 versus 16 discordant pairs; $P = 1.87 \times 10^{-13}$). Excellent Sustainable Habit Building increased from 28.6% (70/245) to 52.2% (128/245; Δ = +23.7 percentage points; 75 versus 17 discordant pairs; $P = 7.29 \times 10^{-10}$). At the opposite end of the ordinal scales, "Absent" behavior-change strategies decreased from 4.1% (10/245) to 0.4% (1/245), while "Poor" habit building decreased from 3.3% (8/245) to 1.2% (3/245).

Counterfactual Coaching was also more frequently judged Realistic and Non-Overwhelming, with top-box ratings increasing from 82.9% (203/245) to 98.8% (242/245; Δ = +15.9 percentage points; 41 versus 2 discordant pairs; $P = 2.15 \times 10^{-10}$). Supportive and Motivating Tone increased from 83.3% (204/245) to 98.4% (241/245; Δ = +15.1 percentage points; 41 versus 4 discordant pairs; $P = 9.33 \times 10^{-9}$).

Not all dimensions differed. Top-box Signal Synthesis ratings were nearly identical between Counterfactual and Generic Coaching (49.0% [120/245] versus 48.6% [119/245]; $P = 1.00$). However, examination of the full ordinal distribution showed that fragmented metric reporting decreased from 19.6% (48/245) in Generic Coaching to 7.8% (19/245) in Counterfactual Coaching, with a corresponding increase in cohesive descriptive summaries (43.3% versus 29.8%).

Lifestyle-coach ratings of Baseline and Wearable Personalization showed a smaller top-box difference. In the evaluations, 54.3% of Counterfactual plans (133/245) and 51.8% of Generic plans (127/245) were rated Highly Tailored ($P = 0.572$). Despite similar top-box rates, bottom-box "Generic" ratings decreased from 4.1% (10/245) for Generic Coaching to 2.4% (6/245) for Counterfactual Coaching.

Easy to Understand was the only dimension with a numerically higher top-box rate for Generic Coaching: 89.8% (220/245) compared with 84.9% (208/245) for Counterfactual Coaching. The difference was not statistically significant (Δ = −4.9 percentage points; 18 versus 30 discordant pairs; $P = 0.111$), but is consistent with a potential trade-off between the simplicity of a single recommendation pathway and the greater numerical density introduced by multiple projected options.

Clinician rubric performance

Clinician evaluations are summarized in Supplementary Table 27. The largest difference was again observed for Contrasts Benefits vs. Status Quo. Counterfactual Coaching achieved the top rating in 92.2% of clinician evaluations (226/245), compared with 7.8% for Generic Coaching (19/245). This represented a +84.5 percentage-point difference, with 213 discordant pairs favoring Counterfactual and 6 favoring Generic ($P = 3.49 \times 10^{-55}$).

Clinicians also rated Counterfactual plans as Highly Tailored to the participant's physiological profile more often than Generic plans (72.2% [177/245] versus 52.2% [128/245]; Δ = +20.0 percentage points; 82 versus 33 discordant pairs; $P = 5.57 \times 10^{-6}$).

Twelve-week Biological Plausibility was rated positively in 86.1% of Counterfactual plans (211/245) compared with 81.6% of Generic plans (200/245), corresponding to a +4.5 percentage-point difference (18 versus 7 discordant pairs; $P = 0.043$). The smaller magnitude of this effect relative to the trajectory-framing and personalization measures indicates that the primary clinician-perceived contribution of the counterfactual system was not a generalized increase across all clinical dimensions, but improved linkage between participant-specific physiology, candidate actions, and projected outcomes.

Cross-disciplinary concordance and divergence

Three evaluation dimensions were shared between lifestyle coaches and clinicians, allowing comparison of how the two expert groups interpreted the same plans.

Both groups strongly favored Counterfactual Coaching for Contrasts Benefits vs. Status Quo, with a pooled difference of +76.1 percentage points ($P < 10^{-98}$). This was the most consistent cross-disciplinary signal and directly reflected the additional future-versus-future information available to the Counterfactual coach.

The groups differed more substantially in their interpretation of Trajectory Control. Clinicians rated both Generic and Counterfactual plans highly on this dimension (93.1% and 96.7%, respectively), reflecting that both conditions provided safe, modifiable, phased recommendations. Lifestyle coaches applied a stricter behavioral interpretation: Generic plans received a top-box rating in 59.4% of evaluations compared

with 97.1% for Counterfactual plans. Coach feedback indicated that the explicit presentation of multiple behavioral options in Counterfactual plans increased perceived autonomy and choice relative to a single prescribed pathway.

The opposite pattern was observed for Baseline and Wearable Personalization. Clinicians gave Counterfactual Coaching a +20.0 percentage-point top-box advantage (72.2% versus 52.2%; $P < 0.001$), whereas lifestyle-coach ratings were more similar (54.3% versus 51.8%; $P = 0.572$). Clinicians appeared to place greater weight on quantitative alignment between the plan and physiological drivers such as Health Age and resting heart rate. Lifestyle coaches noted in free-text responses that physiological tailoring alone did not capture important psychosocial constraints, including work schedules, caregiving responsibilities, and access to exercise resources.

Performance across WEAR-ME Coachable Personas

To evaluate whether the observed preference was restricted to a particular participant phenotype, we stratified results across five prespecified WEAR-ME Coachable Personas (Fig. 5c; Supplementary Table 29).

Among Multi-Risk Challenger participants characterized by low activity, insufficient sleep, elevated BMI, and accelerated Health Age, Counterfactual Coaching achieved its highest side-by-side preference. It was selected as the Best Overall Plan in 80.0% of comparisons (40/50; $P = 2.39 \times 10^{-5}$) and as having the highest likelihood of long-term adherence in 82.0% (41/50; $P = 5.61 \times 10^{-6}$). The largest rubric differences included Trajectory Control (+48.0 percentage points; 100% versus 52%), Behavior-Change Strategies (+32.0 points; 48% versus 16%), Sustainable Habits (+26.0 points; 50% versus 24%), and Realistic and Manageable Routine Fit (+26.0 points; 100% versus 74%). Counterfactual plans more often presented distinct behavioral entry points rather than requiring simultaneous changes across multiple domains.

This subgroup also revealed a calibration limitation. Clinicians rated 12-week Biological Plausibility slightly lower for Counterfactual than Generic plans (76% versus 82%; $\Delta = -6.0$ percentage points). Free-text comments indicated that model-projected Health Age reductions of approximately 4–5 years following modest 12-week changes in activity or sleep appeared overly optimistic in some participant cases.

Among Activity Bottleneck participants with isolated inactivity and preserved sleep, Counterfactual Coaching was preferred in 74.0% of comparisons (37/50; $P = 0.0009$). This subgroup showed the largest persona-specific gains in coach-rated Trajectory Control (+54.0 percentage points; 100% versus 46%) and Behavior-Change Strategies (+34.0 points; 44% versus 10%), as well as a +38.0 percentage-point gain in clinician-rated Baseline Personalization (76% versus 38%). Clinician feedback highlighted that Counterfactual plans more often preserved already-adequate sleep and concentrated recommendations on progressive activity increases.

Among Optimal Baseline participants with high activity, adequate sleep, and a relatively younger Health Age, Counterfactual Coaching was selected as the Best Overall Plan in 72.0% of comparisons (36/50; $P = 0.0026$) and as having the highest likelihood of long-term adherence in 70.0% (35/50; $P = 0.0066$). Counterfactual plans showed higher ratings for Trajectory Control (+34.7 percentage points; 100% versus 65%), Sustainable Habits (+28.0 points; 64% versus 36%), and clinician-rated 12-week Biological Plausibility (+18.0 points; 98% versus 80%). Coaches noted that Generic plans more often prescribed unnecessary increases in activity volume despite already-high baseline activity, whereas Counterfactual plans more often emphasized maintenance and recovery.

Among Balanced Maintainer participants with population-median behavioral profiles, Counterfactual Coaching was selected in 71.1% of comparisons (32/45; $P = 0.0066$) for both Best Overall Plan and anticipated long-term adherence. This subgroup showed the largest persona-specific improvement in Behavior-Change Strategies (+42.2 percentage points; 58% versus 16%) and a +26.7 percentage-point increase in clinician-rated Baseline Personalization (78% versus 51%).

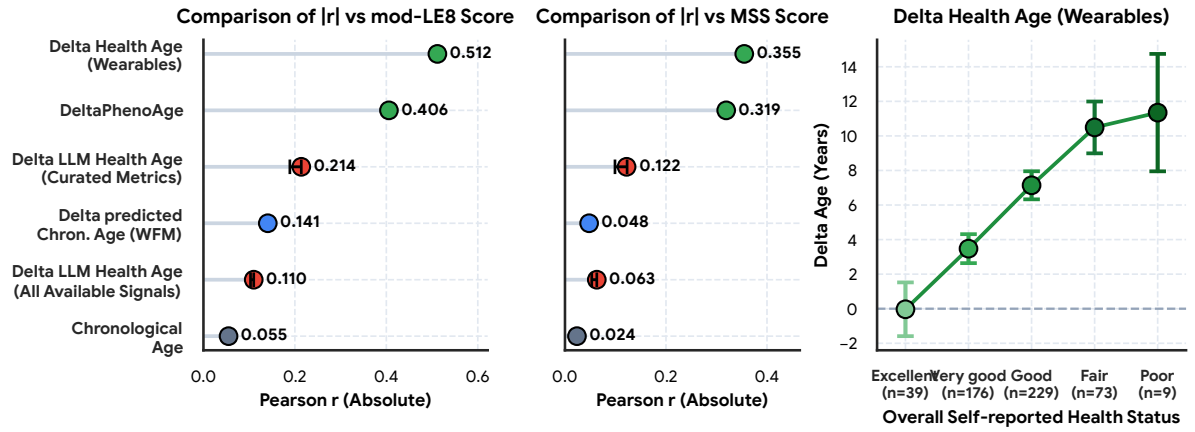
The smallest preference margin occurred among Recovery Bottleneck participants characterized by high activity and insufficient sleep. Counterfactual Coaching was preferred in 60.0% of valid displayed comparisons (30/50) compared with 40.0% for Generic Coaching. Clinicians nevertheless rated Counterfactual plans as more physiologically personalized by +18.0 percentage points (70% versus 52%), in part because the plans more often identified sleep extension as the principal modifiable pathway for already-active participants.

Lifestyle coaches were less favorable in this subgroup, with Baseline and Wearable Personalization ratings 14.0 percentage points lower for Counterfactual Coaching (44% versus 58%). Free-text feedback suggested that some Counterfactual plans continued to include additional step increases for participants already averaging approximately 10,000–14,000 steps per day because their historical peaks exceeded 16,000 steps per day. Coaches viewed these additional activity options as diluting the more appropriate recovery-focused recommendation.

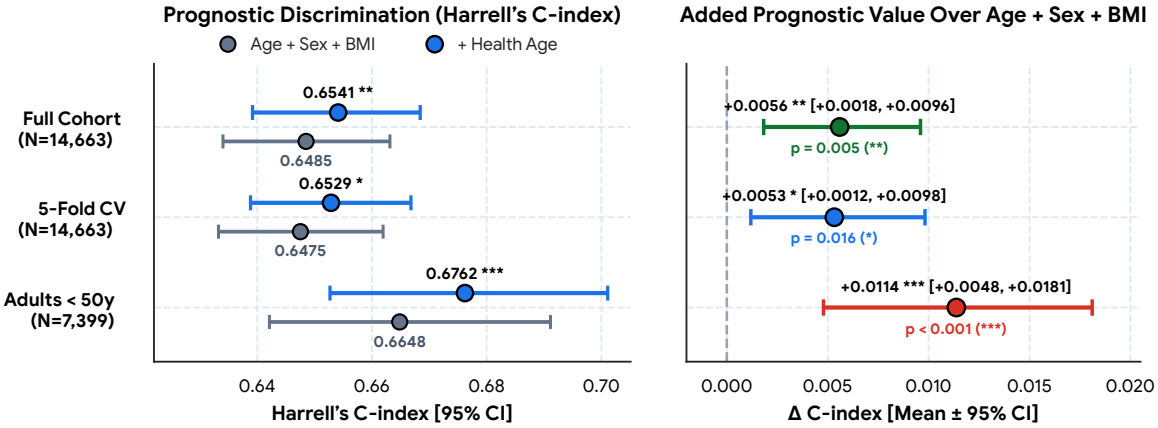
Together, the persona-stratified analyses indicate that the preference for Counterfactual Coaching was not confined to participants with a single behavioral deficit or elevated Health Age. The analysis also identified two recurrent limitations: redundant recommendations in behavioral domains that were already near saturation, and potentially over-optimistic short-horizon Health Age projections in some high-risk participants.

B. Supplementary Figures

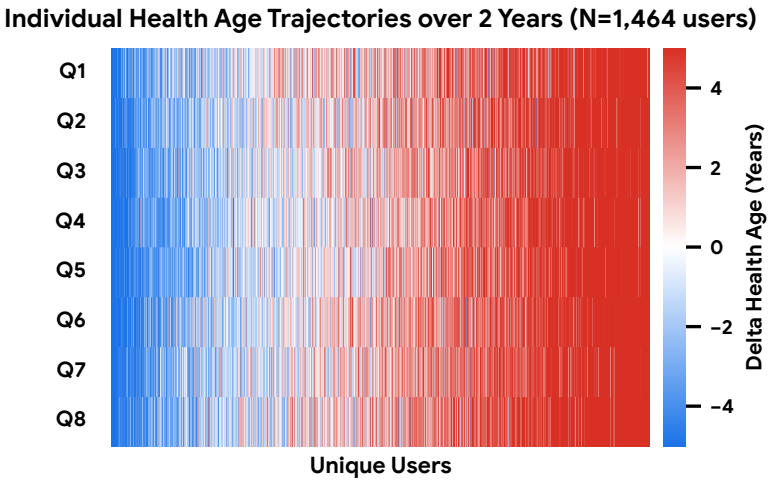
B.1. Coachable Health Age



Supplementary Figure 1 | **Associations Between Coachable Health Age, Cardiometabolic Health, and Self-Reported Health Status.** Absolute Pearson correlation coefficients ($|r|$) comparing Coachable Health Age and alternative age estimates against the modified American Heart Association Life's Essential 8 score (mod-LE8) and Metabolic Syndrome Severity Score (MSSS), alongside wearable-derived Delta Health Age across self-reported health categories, with error bars indicating uncertainty.

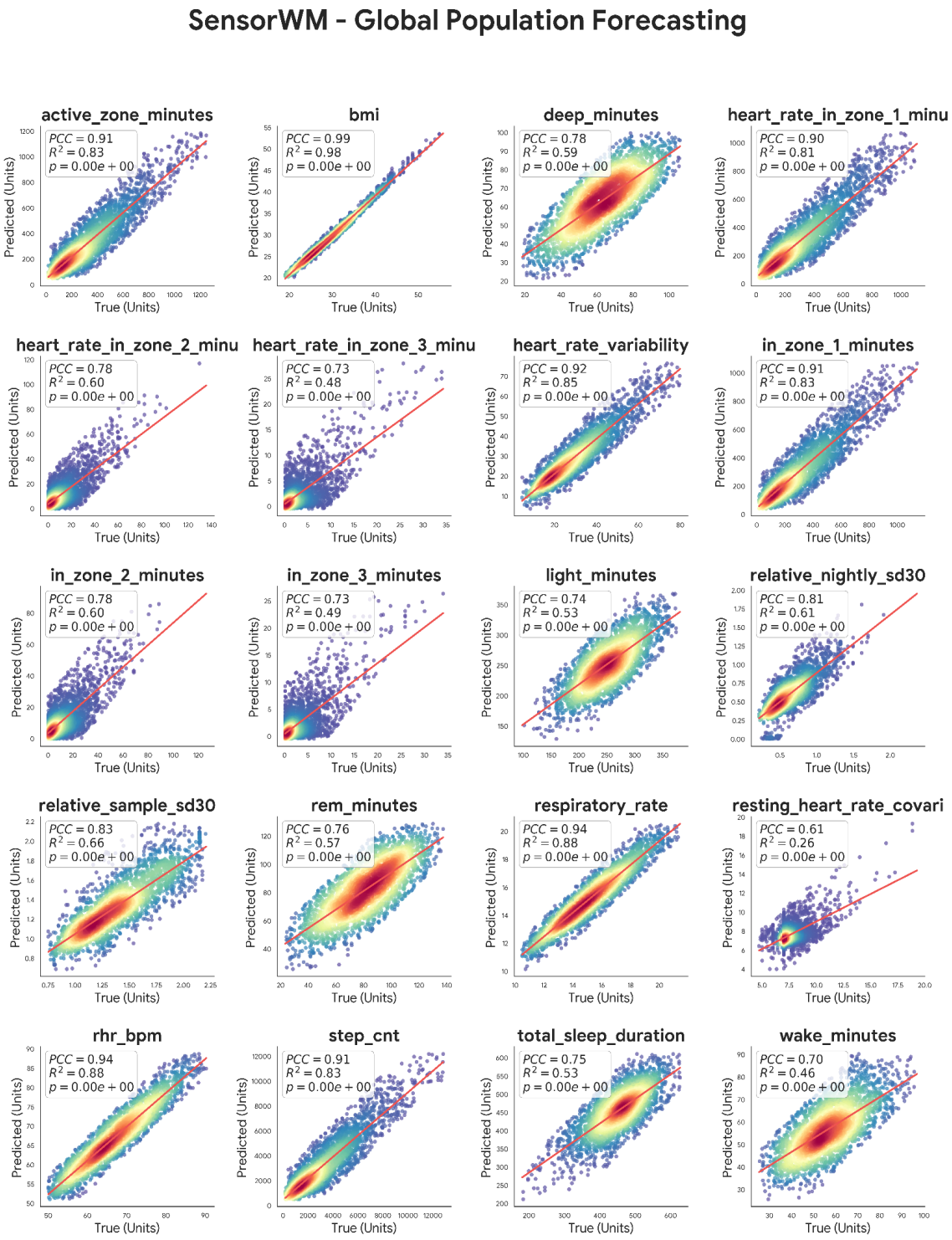


Supplementary Figure 2 | **Health Age provides significant prognostic discrimination beyond age, sex, and BMI in the *All of Us* dataset (N = 14,663).** Harrell's C-index **(left)** and ΔC-index **(right)** (with 95% CIs) show consistent gains in the full cohort (+0.0056**) and 5-fold cross-validation (+0.0053*), with the strongest improvement in adults < 50 years (+0.0114***, N = 7,399). $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p \leq 0.001$.

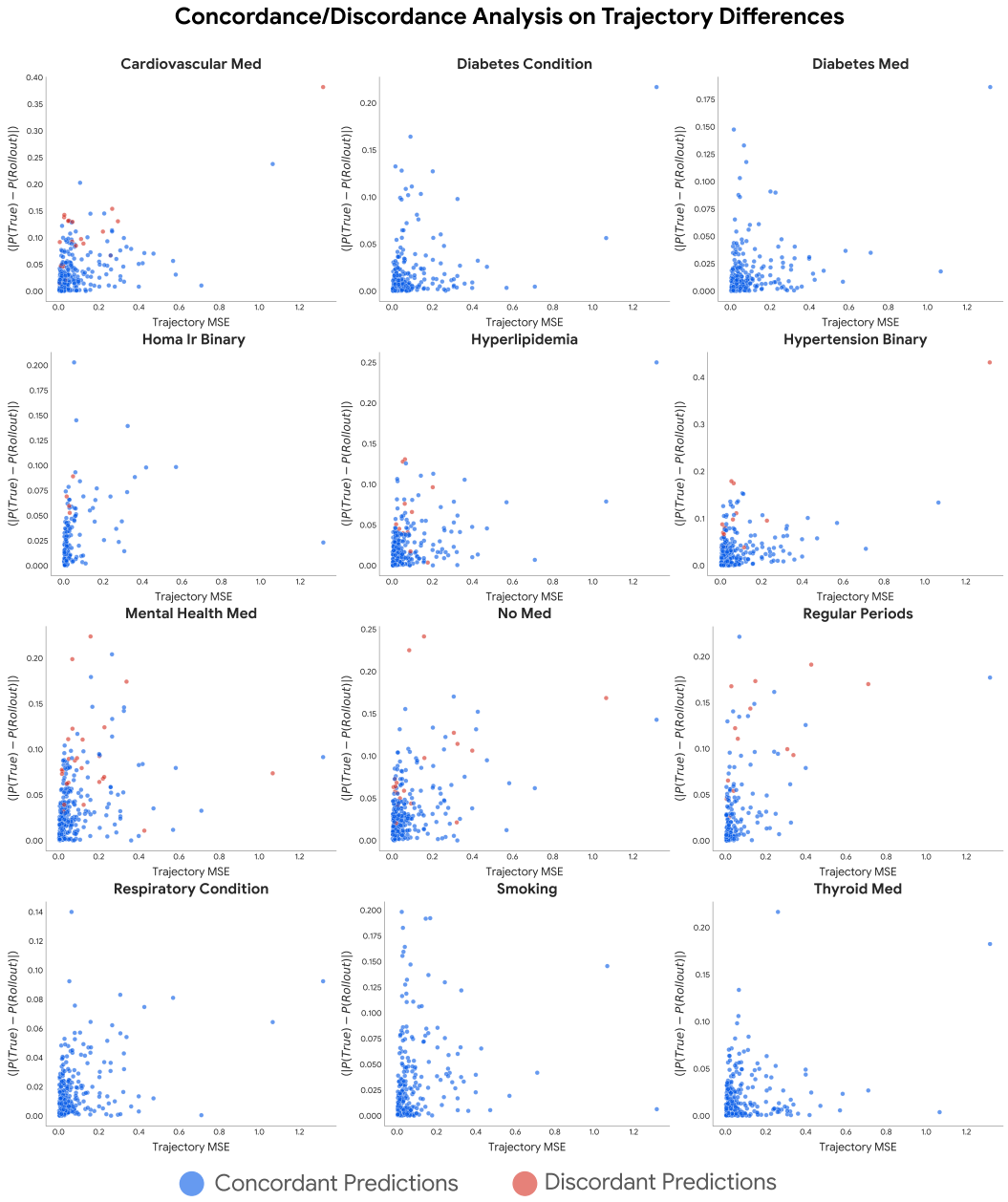


Supplementary Figure 3 | **Individual trajectories of Coachable Health Age over two years.** Heatmap showing longitudinal variation in Health Age Delta (ΔCHA; Coachable Health Age minus chronological age) across 1,464 participants over eight consecutive 3-month windows (Q1–Q8). Each column represents an individual participant, and each row corresponds to a 3-month observation window. Participants are ordered by baseline Health Age Delta (Q1). Colors indicate the magnitude and direction of ΔCHA, with blue denoting a younger physiological health age relative to chronological age (ΔCHA < 0) and red denoting an older physiological health age (ΔCHA > 0). Temporal changes within columns illustrate individual variability in physiological and behavioral health, including trajectories of improvement, deterioration, and relative stability over the two-year observation period.

B.2. Trajectory Prediction with Sensor World Model

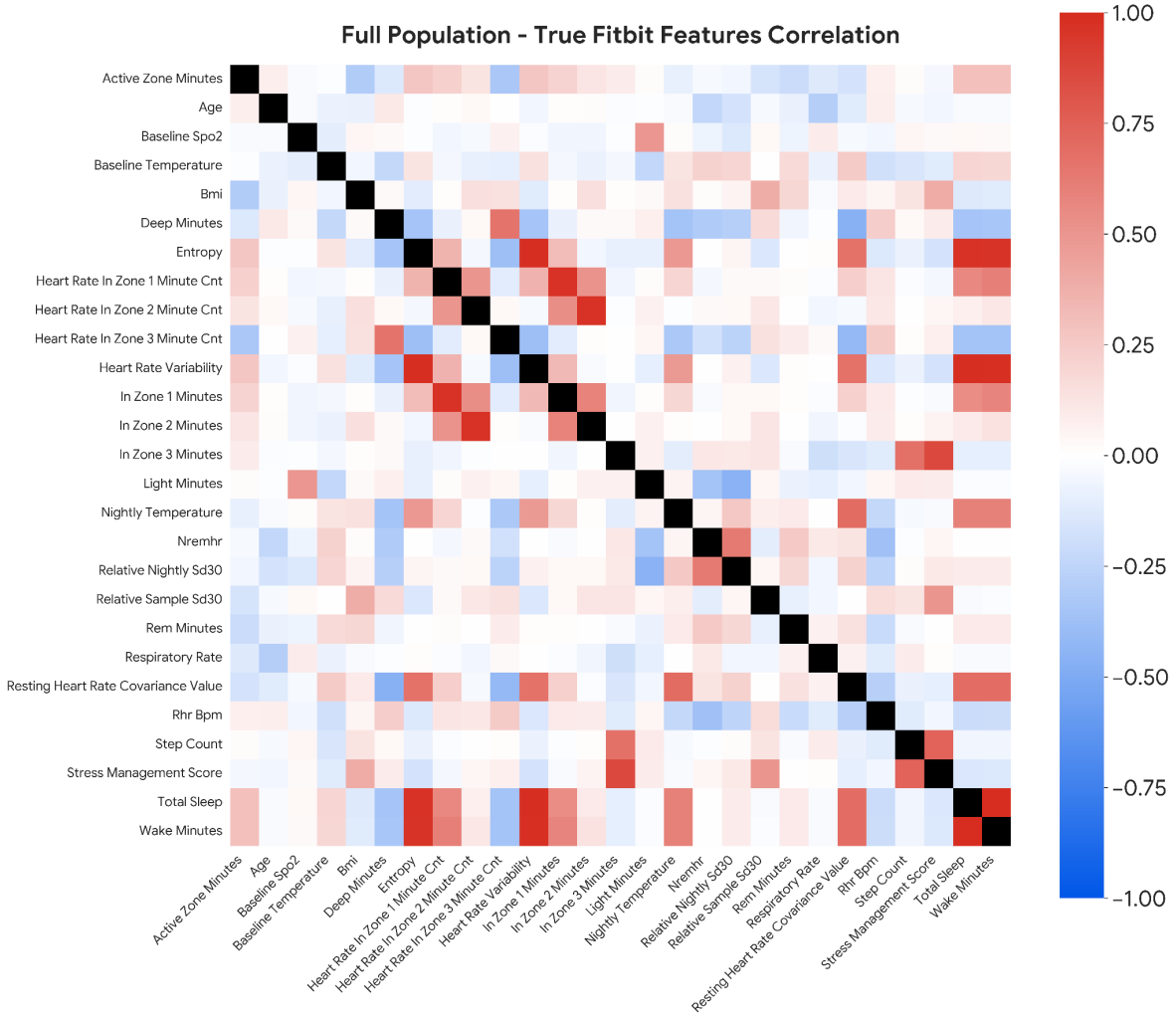


Supplementary Figure 4 | **Population-level forecasting performance of SensorWM.** Comparison of observed and predicted wearable-derived physiological and behavioral measurements across the evaluation cohort. Each panel shows the relationship between ground-truth and model-predicted values for an individual feature, with point density indicated by color and the red diagonal representing perfect agreement. Pearson correlation coefficient (PCC), coefficient of determination (R^2), and associated p-values are reported for each feature.



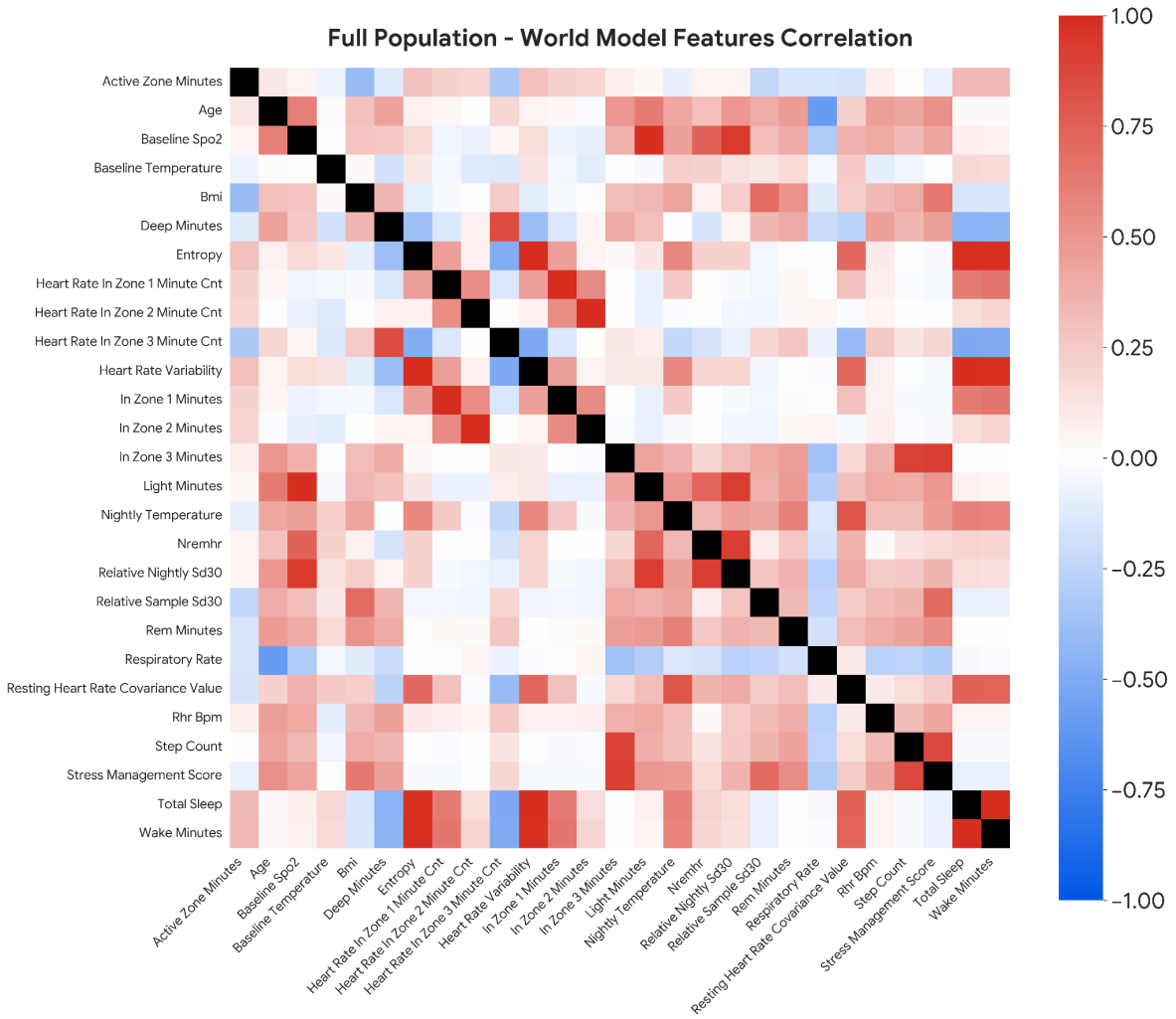
Supplementary Figure 5 | **Trajectory forecasting error and downstream prediction sensitivity across health outcomes.** Relationship between trajectory forecasting error and downstream prediction consistency across clinical and behavioral tasks. Each point corresponds to an individual participant and is plotted according to the trajectory reconstruction error (MSE) between the observed and world-model-generated wearable trajectories (x-axis) and the absolute difference in predicted probability between classifiers trained on true and generated trajectories (y-axis). Blue points indicate concordant classifications, whereas red points indicate discordant classifications. Across most tasks, prediction differences remain small despite variation in trajectory reconstruction error, suggesting that downstream health predictions are relatively robust to moderate forecasting inaccuracies.

(a) Observed Fitbit Features

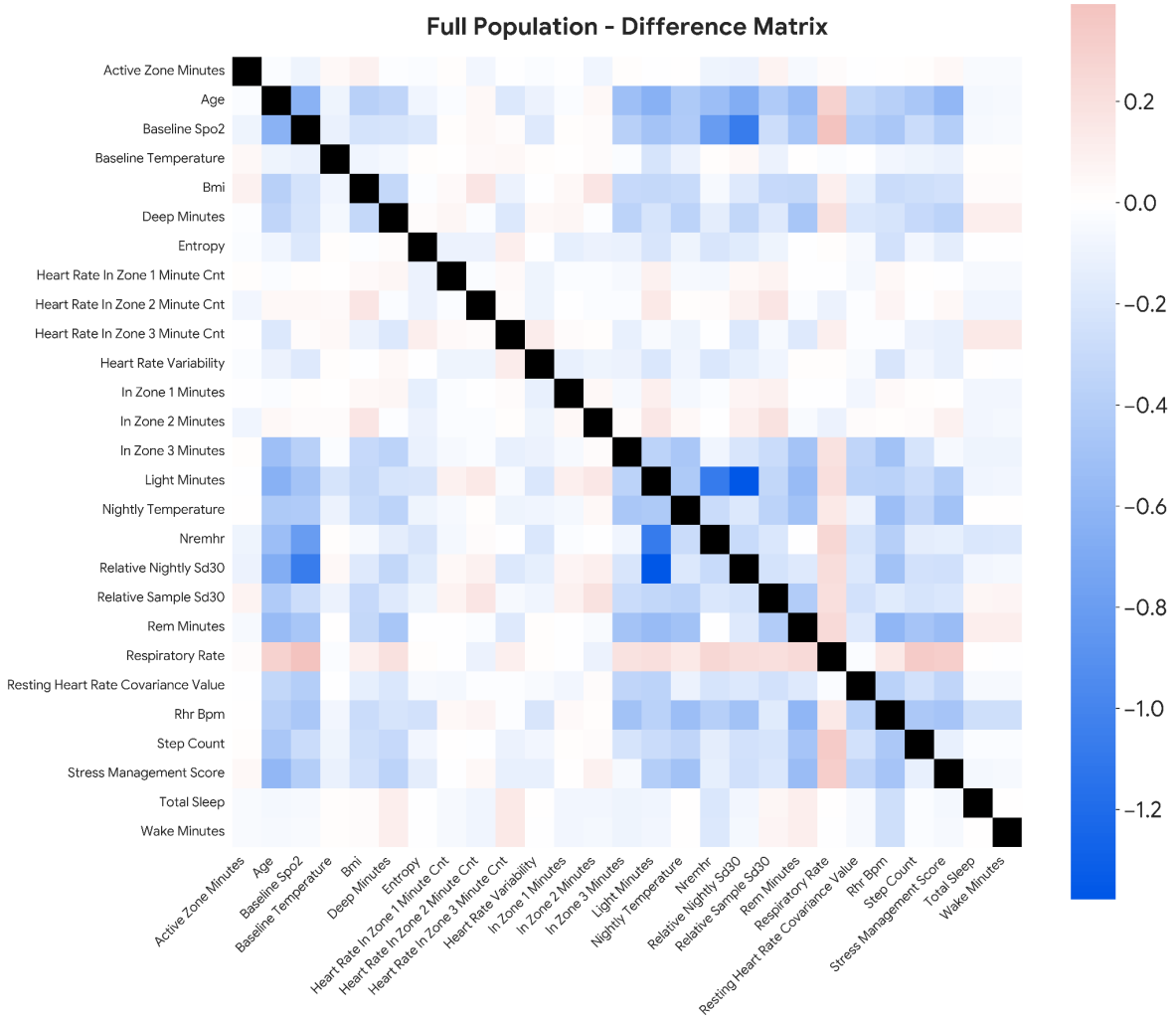


Supplementary Figure 6 | **Pairwise correlation structure of observed and SensorWM-generated wearable features. (a)** Pearson correlation matrix computed from observed Fitbit measurements across the study population. Each cell represents the correlation between two physiological or behavioral features, illustrating the underlying inter-feature dependencies in real wearable data.

(b) SensorWM-Generated Features



Supplementary Figure 6 | **Pairwise correlation structure of observed and SensorWM-generated wearable features (continued). (b)** Pearson correlation matrix computed from SensorWM-generated wearable measurements. Comparison with panel (a) illustrates the extent to which the model preserves positive and negative inter-feature associations observed in real Fitbit data, providing an assessment of the learned joint physiological feature distribution.



Supplementary Figure 7 | **Differences in pairwise correlations between observed and SensorWM-generated Fitbit features.** The matrix shows the difference between pairwise Pearson correlations computed from observed Fitbit features and those generated by SensorWM in the full evaluation population (observed minus generated). Values near zero indicate agreement in feature correlations, whereas positive and negative values indicate stronger correlations in the observed and generated data, respectively. The distribution of differences highlights both preserved relationships and discrepancies in the correlation structure of generated physiological features.

C. Supplementary Tables

C.1. Coachable Health Age

Supplementary Table 1 | **Unadjusted literature hazard ratios, WEAR-ME multicollinearity weights (w_i), and resulting collinearity-adjusted hazard ratios (HR_i^{adj}) for the four core Coachable Health Age domains.**

Core Domain	Knot / Slope Parameter	Reference (HR = 1.0)	Unadjusted HR_i	WEAR-ME Weight (w_i)	Adjusted $HR_i^{adj} = HR_i^{w_i}$
Resting heart rate	Knot at 70 bpm	≤ 60 bpm	1.08	0.7066	1.0559
	Knot at 90 bpm	≤ 60 bpm	1.33	0.7066	1.2232
Daily steps	Per 1,000 steps (cap 9k)	8,000 steps/d	0.88	0.4773	0.9408
Sleep duration	Per hour < 6.0 h	6.0–9.0 h/night	1.20	0.8790	1.1738
	Per hour > 9.0 h	6.0–9.0 h/night	1.34	0.8790	1.2934
Weekly AZM	Knot at 245 min/wk	< 115 min/wk	0.76	0.6950	0.8263
	Knot at 418 min/wk	< 115 min/wk	0.71	0.6950	0.7881
	Plateau ≥ 500 min/wk	< 115 min/wk	0.68	0.6950	0.7648

Supplementary Table 2 | **Comparison of design characteristics across alternative age formulations.** Approaches are summarized according to characteristics relevant to longitudinal wearable health assessment and behavioral guidance. The comparison reflects design considerations motivating Coachable Health Age rather than an empirical comparison of model performance.

Modeling Type	Predictable from Wearables	Explainable	Feature Flexibility	Engageable	Coachable	Movable	Grounded
Chronological Age	✗	✗	✗	✗	✗	✗	✗
Predicted Chronological Age (SensorFM or PPG-FM)	✓	○	✓	○	○	○	✗
PhenoAge	✗	✓	✓	○	○	○	✓
LLM-generated Health Age	✗	✗	✗	✗	✗	✗	✗
Coachable Health Age (wearable features)	✓	✓	✓	✓	✓	✓	✓

✓ indicates that a characteristic is supported by the formulation, ○ indicates limited or indirect support, and ✗ indicates that the characteristic is not supported as defined for the intended longitudinal coaching use case. The categories summarize design considerations and are not empirical performance measures.

Supplementary Table 3 | **Reliability and Consistency of the Wearable Health Age Estimator Across Generation Temperatures.** Comparison of reliability and consistency metrics across different generation temperatures, evaluating the stability and reproducibility of Wearable Health Age estimates under varying sampling conditions.

Input Configuration	Temperature (T)	Average SD (years)	Single-Run Consistency ($ICC(3,1)$)	Average Consistency ($ICC(3,k)$)
Identical Metrics as Health Age	0.1	2.72	0.54	0.78
	0.5	3.54	0.09	0.23
	1.0	3.20	0.11	0.28
All Raw Metrics	0.1	1.35	0.54	0.78
	0.5	2.06	0.30	0.57
	1.0	3.37	0.20	0.42

Supplementary Table 4 | **Bonferroni-Corrected Post-Hoc Comparisons Across Generation Temperatures.** Adjusted p-values from pairwise comparisons across generation temperatures, accounting for multiple hypothesis testing to assess statistically significant differences in estimator performance.

Comparison	**Raw Signals (adj. p)**	**Curated Metrics (adj. p)**
$T_{0.1}$ vs $T_{0.5}$	0.0751	0.0039
$T_{0.5}$ vs $T_{1.0}$	0.0689	0.0024
$T_{0.1}$ vs $T_{1.0}$	1.0000	7.71×10^{-10}

Supplementary Table 5 | **Correlation Between Delta LLM Health Age and Clinical Metrics.** Pearson correlation coefficients (r) quantifying the strength and direction of linear associations between delta LLM health age and clinical metrics, assessing the clinical alignment of wearable-derived LLM health age estimates.

Input Configuration	**Temp (T)**	**Modified LE8 (r)**	**MSSS (r)**
All Available Signals	0.1	-0.110	0.063
	0.5	-0.103	0.061
	1.0	-0.105	0.053
Curated Metrics as Health Age	0.1	**-0.214**	0.122
	0.5	-0.189	0.099
	1.0	-0.192	**0.123**

C.2. Trajectory Prediction with Sensor World Model

Supplementary Table 6 | **State and action features grouped by physiological and behavioral domain.** State variables represent physiological, demographic, and behavioral status variables used by the world model, while action variables represent observable behaviors that may influence future physiological states. Only `step_cnt`, `cardio_minutes`, and `sleep_main_asleep_minute_cnt` are treated as action variables.

Feature	Group	Description	Map
DEM Demographics			
Age	DEM	Chronological age in years.	State
Gender	DEM	Self-reported sex or gender.	State
BMI	DEM	Body mass index.	State
CRD Cardiovascular			
Heart Rate Variability	CRD	Beat-to-beat variability in heart rhythm.	State
Resting Heart Rate BPM	CRD	Resting heart rate in beats per minute.	State
Resting Heart Rate Covariance	CRD	Covariance-based variation in rhr.	State
Heart Rate in Zone 1 Minutes	CRD	Heart rate recorded in zone 1.	State
Heart Rate in Zone 2 Minutes	CRD	Heart rate recorded in zone 2.	State
Heart Rate in Zone 3 Minutes	CRD	Heart rate recorded in zone 3.	State
PLM Cardiopulmonary			
Respiratory Rate	PLM	Average number of breaths per minute.	State
Baseline SpO_2	PLM	Baseline peripheral blood oxygen saturation.	State
SLP Sleep			
Deep Minutes	SLP	Number of minutes spent in deep sleep.	State
Light Minutes	SLP	Number of minutes spent in light sleep.	State
REM Minutes	SLP	Number of minutes spent in REM sleep.	State
Wake Minutes	SLP	Number of wake minutes during sleep.	State
Main Sleep Asleep Minute Count	SLP	Number of minutes asleep.	Action
MTN Motion and Physical Activity			
In Zone 1 Minutes	MTN	Number of activity minutes in zone 1.	State
In Zone 2 Minutes	MTN	Number of activity minutes in zone 2.	State
In Zone 3 Minutes	MTN	Number of activity minutes in zone 3.	State
Step Count	MTN	Step count recorded.	Action
Cardio Minutes	MTN	Number of minutes spent in the cardio-intensity zone.	Action
VAR Regularity and Variability			
Entropy	VAR	Entropy-based measure of temporal regularity and predictability.	State
DeepRMSSD	VAR	Root-mean-square successive-difference statistic computed for the corresponding deep or nightly signal.	State
Relative Nightly SD30	VAR	Relative night-to-night variability estimated over a 30-day window.	State
Relative Sample SD30	VAR	Relative sample-level variability estimated over a 30-day window.	State

Prediction Task	Definition	Type	Source	Metric	N (Pos%)
Health Conditions					
Diabetes Condition	Diagnosis of diabetes condition	CLS	SLF	AUC	1,519 (8.6%)
HOMA-IR	Classification of HOMA-IR ≥ 2.9	CLS	LAB	AUC	775 (31.5%)
Hyperlipidemia	Diagnosis of Hyperlipidemia	CLS	LAB	AUC	1,519 (21.7%)
Hypertension Binary	Diagnosis of hypertension	CLS	SLF	AUC	1,519 (23.6%)
Respiratory Condition	Diagnosis of respiratory condition	CLS	SLF	AUC	1,519 (17.1%)
Behavioral					
Smoking	Smoking behavior	CLS	SLF	AUC	1,495 (8.1%)
Women's Health					
Regular Periods	Self-reported regular menstrual cycles	CLS	SLF	AUC	871 (41.4%)
Medication					
Cardiovascular Medication	Use of cardiovascular medication	CLS	SLF	AUC	1,654 (33.0%)
Diabetes Medication	Use of diabetes medication	CLS	SLF	AUC	1,654 (8.7%)
Mental Health Medication	Use of depression/anxiety medication	CLS	SLF	AUC	1,654 (27.4%)
No Medication	No regular medication use in daily life	CLS	SLF	AUC	1,654 (38.8%)
Thyroid Medication	Use of thyroid medication	CLS	SLF	AUC	1,654 (9.9%)

Supplementary Table 7 | **Downstream prediction tasks grouped into categories.** Source indicates whether labels originate from self-reported questionnaires (SLF) or laboratory measurements (LAB). *N* denotes the number of users with valid labels and Pos% denotes positive prevalence among valid users.

Supplementary Table 8 | **Demographic characteristics of the pre-training cohort.** Distribution of age, body mass index (BMI), and sex among participants in the Fitbit pre-training cohort.

Category	N	%
Age		
<40	54,050	55.8
40–59	27,514	28.3
60+	15,393	15.9
BMI		
<25	47,947	49.5
25–30	27,481	28.3
>30	21,505	22.2
Unknown	24	0.0
Sex		
Male	49,294	50.8
Female	47,651	49.2
Unknown	12	0.0
Total Users	**96,957**	**100.0**

Supplementary Table 9 | **Demographic characteristics of the WEAR-ME evaluation cohort.** Distribution of age, body mass index (BMI), and sex among participants with available downstream clinical and behavioral labels.

Category	N	%
Age		
<40	398	24.1
40–59	937	56.7
60+	319	19.3
BMI		
<25	478	28.9
25–30	499	30.2
>30	677	40.9
Sex		
Male	659	39.8
Female	993	60.0
Unknown	2	0.1
Total Users	**1,654**	**100.0**

Supplementary Table 10 | **Downstream task splits.** Each row is a prediction task, and each split reports the number of valid samples, the number of positive samples, and the positive prevalence among valid samples.

Prediction Task	Train			Val			Test		
	N	Pos	Pos%	N	Pos	Pos%	N	Pos	Pos%
Health Conditions									
Diabetes Condition	899	85	9.5%	315	20	6.3%	305	26	8.5%
HOMA-IR Binary	470	150	31.9%	164	56	34.1%	141	38	27.0%
Hyperlipidemia	899	194	21.6%	315	76	24.1%	305	60	19.7%
Hypertension Binary	899	214	23.8%	315	82	26.0%	305	63	20.7%
Respiratory Condition	899	147	16.4%	315	56	17.8%	305	57	18.7%
Behavioral									
Smoking	884	69	7.8%	311	30	9.6%	300	22	7.3%
Women's Health									
Regular Periods	516	216	41.9%	188	79	42.0%	167	66	39.5%
Medication									
Cardiovascular Medication	992	326	32.9%	331	118	35.6%	331	102	30.8%
Diabetes Medication	992	85	8.6%	331	27	8.2%	331	32	9.7%
Mental Health Medication	992	257	25.9%	331	99	29.9%	331	97	29.3%
No Medication	992	388	39.1%	331	123	37.2%	331	131	39.6%
Thyroid Medication	992	92	9.3%	331	44	13.3%	331	28	8.5%

Supplementary Table 11 | **Forecasting Performance of the Sensor World Model (R^2).** Forecasting accuracy is evaluated using the coefficient of determination (R^2), with results reported as mean ± standard error across evaluation runs.

Prediction Task	Type	Metric	Feature Source					
			1 Week Mean	2 Week Mean	4 Week Mean	LOCF	TimesFM	SensorWM
Forecasting								
active_zone_minutes	FOR	r^2	0.4937 ± 0.0025	0.6465 ± 0.0020	0.7166 ± 0.0017	0.7129 ± 0.0017	0.6355 ± 0.0047	**0.8260 ± 0.0012**
bmi	FOR	r^2	**1.0000 ± 0.0000**	**1.0000 ± 0.0000**	**1.0000 ± 0.0000**	**1.0000 ± 0.0000**	0.9352 ± 0.0010	0.9835 ± 0.0001
deep_minutes	FOR	r^2	-0.4998 ± 0.0000	0.1603 ± 0.0023	0.3917 ± 0.0026	0.3302 ± 0.0055	0.3421 ± 0.0062	**0.5929 ± 0.0023**
heart_rate_in_zone_1_minute_cnt	FOR	r^2	0.4006 ± 0.0026	0.5808 ± 0.0022	0.6699 ± 0.0019	0.6681 ± 0.0043	0.6215 ± 0.0048	**0.8143 ± 0.0012**
heart_rate_in_zone_2_minute_cnt	FOR	r^2	0.1016 ± 0.0020	0.4582 ± 0.0025	0.5392 ± 0.0023	0.5981 ± 0.0036	0.5468 ± 0.0054	**0.6015 ± 0.0023**
heart_rate_in_zone_3_minute_cnt	FOR	r^2	-0.3634 ± 0.0000	-0.0090 ± 0.0000	0.1021 ± 0.0020	0.3872 ± 0.0062	0.3518 ± 0.0062	**0.4789 ± 0.0037**
heart_rate_variability	FOR	r^2	0.2998 ± 0.0027	0.5562 ± 0.0023	0.6629 ± 0.0019	0.6818 ± 0.0018	0.6052 ± 0.0050	**0.8545 ± 0.0010**
in_zone_1_minutes	FOR	r^2	0.4927 ± 0.0025	0.6423 ± 0.0020	0.7136 ± 0.0017	0.7085 ± 0.0017	0.6443 ± 0.0046	**0.8256 ± 0.0012**
in_zone_2_minutes	FOR	r^2	0.1311 ± 0.0022	0.4895 ± 0.0025	0.5563 ± 0.0023	**0.6262 ± 0.0021**	0.4519 ± 0.0060	0.5989 ± 0.0023
in_zone_3_minutes	FOR	r^2	-0.4021 ± 0.0000	-0.0102 ± 0.0000	0.1048 ± 0.0020	-0.0651 ± 0.0000	0.3804 ± 0.0062	**0.4924 ± 0.0037**
light_minutes	FOR	r^2	-1.5211 ± 0.0000	-0.3540 ± 0.0000	0.1365 ± 0.0022	0.2113 ± 0.0025	0.2405 ± 0.0060	**0.5321 ± 0.0025**
relative_nightly_sd30	FOR	r^2	0.4291 ± 0.0026	0.4546 ± 0.0026	0.5277 ± 0.0024	**0.7147 ± 0.0017**	0.5540 ± 0.0054	0.6143 ± 0.0036
relative_sample_sd30	FOR	r^2	0.6871 ± 0.0018	0.7389 ± 0.0016	0.7985 ± 0.0013	**0.9108 ± 0.0010**	0.8460 ± 0.0023	0.6588 ± 0.0032
rem_minutes	FOR	r^2	-0.7262 ± 0.0000	0.0768 ± 0.0018	0.3648 ± 0.0027	0.3050 ± 0.0048	0.3095 ± 0.0062	**0.5726 ± 0.0024**
respiratory_rate	FOR	r^2	-1.9790 ± 0.0000	-0.3847 ± 0.0000	0.2893 ± 0.0027	0.6225 ± 0.0059	0.5034 ± 0.0057	**0.8835 ± 0.0008**
resting_heart_rate_covariance_value	FOR	r^2	-14.5124 ± 0.0000	-3.6321 ± 0.0000	-0.8187 ± 0.0000	-0.1515 ± 0.0000	**0.3760 ± 0.0062**	0.2618 ± 0.0028
rhr_bpm	FOR	r^2	-0.7103 ± 0.0000	0.0667 ± 0.0017	0.4682 ± 0.0025	0.7049 ± 0.0017	0.6864 ± 0.0042	**0.8769 ± 0.0008**
step_cnt	FOR	r^2	0.4904 ± 0.0025	0.6454 ± 0.0020	0.7151 ± 0.0017	0.7123 ± 0.0017	0.6341 ± 0.0047	**0.8252 ± 0.0012**
total_sleep_duration	FOR	r^2	-1.8114 ± 0.0000	-0.4606 ± 0.0000	0.0903 ± 0.0019	0.1960 ± 0.0025	0.4182 ± 0.0061	**0.5299 ± 0.0025**
wake_minutes	FOR	r^2	-0.9326 ± 0.0000	-0.1221 ± 0.0000	0.2334 ± 0.0026	0.1840 ± 0.0000	0.1867 ± 0.0057	**0.4599 ± 0.0027**

Supplementary Table 12 | **Forecasting Performance of the Sensor World Model (r).** Forecasting accuracy is evaluated using the Pearson correlation coefficient (r), with results reported as mean ± standard error across evaluation runs.

Prediction Task	Type	Metric	Feature Source					
			1 Week Mean	2 Week Mean	4 Week Mean	LOCF	TimesFM	SensorWM
Forecasting								
active_zone_minutes	FOR	r	0.7791 ± 0.0022	0.8287 ± 0.0019	0.8565 ± 0.0018	0.8583 ± 0.0018	0.7985 ± 0.0049	**0.9098 ± 0.0015**
bmi	FOR	r	**1.0000 ± 0.0000**	**1.0000 ± 0.0000**	**1.0000 ± 0.0000**	**1.0000 ± 0.0000**	0.9693 ± 0.0020	0.9942 ± 0.0004
deep_minutes	FOR	r	0.5362 ± 0.0029	0.6379 ± 0.0027	0.6970 ± 0.0025	0.6713 ± 0.0026	0.5966 ± 0.0065	**0.7769 ± 0.0023**
heart_rate_in_zone_1_minute_cnt	FOR	r	0.7442 ± 0.0023	0.7997 ± 0.0021	0.8332 ± 0.0019	0.8370 ± 0.0019	0.7913 ± 0.0050	**0.9034 ± 0.0016**
heart_rate_in_zone_2_minute_cnt	FOR	r	0.6831 ± 0.0025	0.7336 ± 0.0024	0.7515 ± 0.0023	**0.7939 ± 0.0021**	0.7543 ± 0.0053	0.7842 ± 0.0023
heart_rate_in_zone_3_minute_cnt	FOR	r	0.2592 ± 0.0034	0.3598 ± 0.0032	0.4205 ± 0.0031	0.5106 ± 0.0030	0.6116 ± 0.0064	**0.7292 ± 0.0035**
heart_rate_variability	FOR	r	0.6864 ± 0.0025	0.7792 ± 0.0022	0.8258 ± 0.0020	0.8418 ± 0.0019	0.7816 ± 0.0051	**0.9246 ± 0.0014**
in_zone_1_minutes	FOR	r	0.7764 ± 0.0022	0.8267 ± 0.0020	0.8551 ± 0.0018	0.8563 ± 0.0018	0.8056 ± 0.0048	**0.9095 ± 0.0015**
in_zone_2_minutes	FOR	r	0.6957 ± 0.0025	0.7479 ± 0.0023	0.7643 ± 0.0022	**0.8086 ± 0.0020**	0.6819 ± 0.0059	0.7829 ± 0.0023
in_zone_3_minutes	FOR	r	0.2620 ± 0.0034	0.3689 ± 0.0032	0.4380 ± 0.0031	0.5377 ± 0.0029	0.6371 ± 0.0062	**0.7343 ± 0.0035**
light_minutes	FOR	r	0.3773 ± 0.0032	0.5065 ± 0.0030	0.6082 ± 0.0028	0.6315 ± 0.0027	0.5484 ± 0.0068	**0.7443 ± 0.0024**
relative_nightly_sd30	FOR	r	0.6839 ± 0.0025	0.7001 ± 0.0025	0.7449 ± 0.0023	**0.8533 ± 0.0018**	0.7523 ± 0.0053	0.8067 ± 0.0036
relative_sample_sd30	FOR	r	0.8458 ± 0.0019	0.8707 ± 0.0017	0.8998 ± 0.0015	**0.9555 ± 0.0010**	0.9256 ± 0.0031	0.8278 ± 0.0033
rem_minutes	FOR	r	0.4504 ± 0.0031	0.5856 ± 0.0028	0.6716 ± 0.0026	0.6577 ± 0.0026	0.5743 ± 0.0066	**0.7642 ± 0.0023**
respiratory_rate	FOR	r	0.4182 ± 0.0032	0.5943 ± 0.0028	0.7340 ± 0.0024	0.8337 ± 0.0019	0.7403 ± 0.0054	**0.9405 ± 0.0013**
resting_heart_rate_covariance_value	FOR	r	0.1161 ± 0.0034	0.2317 ± 0.0034	0.3665 ± 0.0032	0.4788 ± 0.0030	**0.6198 ± 0.0064**	0.6114 ± 0.0029
rhr_bpm	FOR	r	0.5315 ± 0.0029	0.6712 ± 0.0026	0.7773 ± 0.0022	0.8670 ± 0.0017	0.8303 ± 0.0045	**0.9366 ± 0.0012**
step_cnt	FOR	r	0.7783 ± 0.0022	0.8279 ± 0.0019	0.8556 ± 0.0018	0.8580 ± 0.0018	0.7978 ± 0.0049	**0.9093 ± 0.0015**
total_sleep_duration	FOR	r	0.3925 ± 0.0032	0.5122 ± 0.0030	0.6076 ± 0.0028	0.6319 ± 0.0027	0.6594 ± 0.0061	**0.7461 ± 0.0024**
wake_minutes	FOR	r	0.4260 ± 0.0031	0.5294 ± 0.0029	0.6173 ± 0.0027	0.6033 ± 0.0028	0.4861 ± 0.0071	**0.7000 ± 0.0026**

Supplementary Table 13 | **Trajectory Based versus Static Downstream Classification Performance.** Classification performance comparing downstream prediction using true Fitbit features against Trajectory Transformer and LeWorldModel. Values are AUROC with 95% confidence intervals.

Prediction Task	Type	Metric	Feature Source			
			N (Pos%)	True Fitbit Features	Trajectory Transformer	LEWorldModel
Health Conditions						
Diabetes Condition	CLS	AUC	305 (8.5%)	0.723 [0.64–0.80]	0.713 [0.60–0.85]	**0.724 [0.63–0.83]**
HOMA-IR Binary	CLS	AUC	141 (27.0%)	0.780 [0.70–0.86]	0.719 [0.64–0.82]	**0.760 [0.67–0.85]**
Hyperlipidemia	CLS	AUC	305 (19.7%)	0.691 [0.62–0.76]	0.645 [0.58–0.73]	**0.692 [0.61–0.76]**
Hypertension Binary	CLS	AUC	305 (20.7%)	0.796 [0.74–0.86]	**0.793 [0.73–0.86]**	**0.793 [0.73–0.85]**
Respiratory Condition	CLS	AUC	305 (18.7%)	0.629 [0.57–0.71]	0.634 [0.56–0.73]	**0.660 [0.58–0.72]**
Behavioral						
Smoking	CLS	AUC	300 (7.3%)	0.784 [0.69–0.87]	0.698 [0.57–0.81]	**0.774 [0.69–0.85]**
Women's Health						
Regular Periods	CLS	AUC	167 (39.5%)	0.852 [0.79–0.91]	**0.840 [0.78–0.89]**	0.834 [0.77–0.90]
Medication						
Cardiovascular Medication	CLS	AUC	331 (30.8%)	0.785 [0.75–0.84]	0.766 [0.72–0.83]	**0.775 [0.73–0.83]**
Diabetes Medication	CLS	AUC	331 (9.7%)	0.719 [0.63–0.80]	0.692 [0.58–0.77]	**0.730 [0.64–0.80]**
Mental Health Medication	CLS	AUC	331 (29.3%)	0.631 [0.57–0.69]	**0.671 [0.61–0.73]**	0.616 [0.56–0.68]
No Medication	CLS	AUC	331 (39.6%)	0.710 [0.64–0.76]	0.683 [0.62–0.74]	**0.695 [0.64–0.75]**
Thyroid Medication	CLS	AUC	331 (8.5%)	0.750 [0.65–0.84]	0.643 [0.54–0.73]	**0.716 [0.62–0.80]**

Supplementary Table 14 | **Trajectory Based versus Static Downstream Classification Performance across Data Scales.** Classification performance comparing downstream prediction using true Fitbit features against LEWorldModel representations fine-tuned across different scale baselines (1k, 10k, and 100k). Values are AUROC with 95% confidence intervals.

Prediction Task	Type	Metric	Feature Source				
			N (Pos%)	True Fitbit Features	1k	10k	100k
Health Conditions							
Diabetes Condition	CLS	AUC	305 (8.5%)	0.723 [0.64–0.80]	**0.736 [0.66–0.81]**	0.706 [0.61–0.80]	0.724 [0.63–0.83]
HOMA-IR Binary	CLS	AUC	141 (27.0%)	0.780 [0.70–0.86]	0.699 [0.61–0.81]	0.710 [0.61–0.80]	**0.760 [0.67–0.85]**
Hyperlipidemia	CLS	AUC	305 (19.7%)	0.691 [0.62–0.76]	0.673 [0.61–0.75]	0.666 [0.60–0.75]	**0.692 [0.61–0.76]**
Hypertension Binary	CLS	AUC	305 (20.7%)	0.796 [0.74–0.86]	**0.807 [0.75–0.87]**	0.797 [0.74–0.85]	0.793 [0.73–0.85]
Respiratory Condition	CLS	AUC	305 (18.7%)	0.629 [0.57–0.71]	0.619 [0.56–0.70]	0.579 [0.51–0.65]	**0.660 [0.58–0.72]**
Behavioral							
Smoking	CLS	AUC	300 (7.3%)	0.784 [0.69–0.87]	0.727 [0.65–0.81]	0.670 [0.57–0.76]	**0.774 [0.69–0.85]**
Women's Health							
Regular Periods	CLS	AUC	167 (39.5%)	0.852 [0.79–0.91]	0.815 [0.76–0.87]	0.812 [0.75–0.87]	**0.834 [0.77–0.90]**
Medication							
Cardiovascular Medication	CLS	AUC	331 (30.8%)	0.785 [0.75–0.84]	0.768 [0.73–0.83]	0.769 [0.72–0.84]	**0.775 [0.73–0.83]**
Diabetes Medication	CLS	AUC	331 (9.7%)	0.719 [0.63–0.80]	**0.736 [0.64–0.81]**	0.734 [0.63–0.82]	0.730 [0.64–0.80]
Mental Health Medication	CLS	AUC	331 (29.3%)	0.631 [0.57–0.69]	**0.630 [0.58–0.68]**	0.627 [0.57–0.68]	0.616 [0.56–0.68]
No Medication	CLS	AUC	331 (39.6%)	0.710 [0.64–0.76]	0.678 [0.61–0.73]	0.687 [0.63–0.74]	**0.695 [0.64–0.75]**
Thyroid Medication	CLS	AUC	331 (8.5%)	0.750 [0.65–0.84]	0.669 [0.57–0.76]	0.701 [0.62–0.77]	**0.716 [0.62–0.80]**

Supplementary Table 15 | **Trajectory Based versus Static Downstream Classification Performance.** Classification performance comparing downstream prediction using true Fitbit features, world-model-generated Fitbit features, SensorFM, and first-week-only Fitbit features. Values are AUROC with 95% confidence intervals.

Prediction Task	Type	Metric	Feature Source				
			N (Pos%)	True Fitbit Features	Sensor World Model Fitbit Features	SensorFM	First Week Fitbit Features
Health Conditions							
Diabetes Condition	CLS	AUC	305 (8.5%)	0.723 [0.64–0.80]	0.724 [0.63–0.83]	0.643 [0.52–0.75]	0.655 [0.54–0.77]
HOMA-IR Binary	CLS	AUC	141 (27.0%)	0.780 [0.70–0.86]	0.760 [0.67–0.85]	0.774 [0.69–0.85]	0.717 [0.61–0.82]
Hyperlipidemia	CLS	AUC	305 (19.7%)	0.691 [0.62–0.76]	0.692 [0.61–0.76]	0.614 [0.53–0.69]	0.535 [0.45–0.63]
Hypertension Binary	CLS	AUC	305 (20.7%)	0.796 [0.74–0.86]	0.793 [0.73–0.85]	0.753 [0.68–0.82]	0.766 [0.70–0.83]
Respiratory Condition	CLS	AUC	305 (18.7%)	0.629 [0.57–0.71]	0.660 [0.58–0.72]	0.658 [0.58–0.73]	0.535 [0.45–0.62]
Behavioral							
Smoking	CLS	AUC	300 (7.3%)	0.784 [0.69–0.87]	0.774 [0.69–0.85]	0.666 [0.55–0.77]	0.595 [0.47–0.71]
Women's Health							
Regular Periods	CLS	AUC	167 (39.5%)	0.852 [0.79–0.91]	0.834 [0.77–0.90]	0.766 [0.69–0.84]	0.745 [0.67–0.82]
Medication							
Cardiovascular Medication	CLS	AUC	331 (30.8%)	0.785 [0.75–0.84]	0.775 [0.73–0.83]	0.744 [0.68–0.80]	0.704 [0.64–0.76]
Diabetes Medication	CLS	AUC	331 (9.7%)	0.719 [0.63–0.80]	0.730 [0.64–0.80]	0.624 [0.51–0.74]	0.634 [0.52–0.75]
Mental Health Medication	CLS	AUC	331 (29.3%)	0.631 [0.57–0.69]	0.616 [0.56–0.68]	0.784 [0.73–0.83]	0.614 [0.54–0.68]
No Medication	CLS	AUC	331 (39.6%)	0.710 [0.64–0.76]	0.695 [0.64–0.75]	0.696 [0.64–0.75]	0.647 [0.58–0.71]
Thyroid Medication	CLS	AUC	331 (8.5%)	0.750 [0.65–0.84]	0.716 [0.62–0.80]	0.583 [0.48–0.67]	0.602 [0.49–0.71]

Supplementary Table 16 | **Classification Performance of Fitbit and SensorWM Features on Entire Cohort.** Classification performance comparing downstream prediction using true Fitbit features and world-model-generated Fitbit features. Values are AUROC with 95% confidence intervals. Statistical significance between AUROC estimates was assessed using DeLong's test.

Prediction Task	Type	Metric	Feature Source			
			N (Pos%)	True Fitbit Features	World Model Fitbit Features	DeLong p-val
Health Conditions						
Diabetes Condition	CLS	AUC	305 (8.5%)	0.723 [0.64–0.80]	0.724 [0.63–0.83]	0.7037
HOMA-IR Binary	CLS	AUC	141 (27.0%)	0.780 [0.70–0.86]	0.760 [0.67–0.85]	0.2782
Hyperlipidemia	CLS	AUC	305 (19.7%)	0.691 [0.62–0.76]	0.692 [0.61–0.76]	0.9234
Hypertension Binary	CLS	AUC	305 (20.7%)	0.796 [0.74–0.86]	0.793 [0.73–0.85]	0.7784
Respiratory Condition	CLS	AUC	305 (18.7%)	0.629 [0.57–0.71]	0.660 [0.58–0.72]	0.6494
Behavioral						
Smoking	CLS	AUC	300 (7.3%)	0.784 [0.69–0.87]	0.774 [0.69–0.85]	0.2013
Women's Health						
Regular Periods	CLS	AUC	167 (39.5%)	0.852 [0.79–0.91]	0.834 [0.77–0.90]	0.0328
Medication						
Cardiovascular Medication	CLS	AUC	331 (30.8%)	0.785 [0.75–0.84]	0.775 [0.73–0.83]	0.1427
Diabetes Medication	CLS	AUC	331 (9.7%)	0.719 [0.63–0.80]	0.730 [0.64–0.80]	0.0639
Mental Health Medication	CLS	AUC	331 (29.3%)	0.631 [0.57–0.69]	0.616 [0.56–0.68]	0.2299
No Medication	CLS	AUC	331 (39.6%)	0.710 [0.64–0.76]	0.695 [0.64–0.75]	0.1244
Thyroid Medication	CLS	AUC	331 (8.5%)	0.750 [0.65–0.84]	0.716 [0.62–0.80]	0.0332

Supplementary Table 17 | **Classification Performance of Fitbit and SensorWM Features on Users Ages 18-39.** Classification performance comparing downstream prediction using true Fitbit features and world-model-generated Fitbit features for the age bucket 18–39. Values are AUROC with 95% confidence intervals. Statistical significance between AUROC estimates was assessed using DeLong's test.

Prediction Task	Type	Metric	Feature Source			
			N (Pos%)	True Fitbit Features	World Model Fitbit Features	DeLong p-val
Health Conditions						
Diabetes Condition	CLS	AUC	74 (2.7%)	0.875 [0.71–1.00]	0.792 [0.62–0.93]	0.0006
HOMA-IR Binary	CLS	AUC	33 (27.3%)	0.764 [0.59–0.89]	0.755 [0.57–0.89]	0.8322
Hyperlipidemia	CLS	AUC	74 (8.1%)	0.424 [0.18–0.74]	0.414 [0.16–0.75]	0.8301
Hypertension Binary	CLS	AUC	74 (8.1%)	0.787 [0.51–0.98]	0.767 [0.46–0.98]	0.3932
Respiratory Condition	CLS	AUC	74 (13.5%)	0.608 [0.47–0.79]	0.564 [0.41–0.77]	0.2493
Behavioral						
Smoking	CLS	AUC	71 (11.3%)	0.837 [0.72–0.97]	0.669 [0.51–0.83]	0.0127
Women's Health						
Regular Periods	CLS	AUC	36 (69.4%)	0.538 [0.36–0.76]	0.549 [0.38–0.76]	0.7467
Medication						
Cardiovascular Medication	CLS	AUC	81 (13.6%)	0.673 [0.52–0.83]	0.604 [0.45–0.78]	0.0008
Diabetes Medication	CLS	AUC	81 (3.7%)	0.893 [0.81–0.96]	0.821 [0.74–0.91]	0.0129
Mental Health Medication	CLS	AUC	81 (33.3%)	0.664 [0.55–0.78]	0.623 [0.52–0.73]	0.1587
No Medication	CLS	AUC	81 (49.4%)	0.751 [0.66–0.85]	0.746 [0.66–0.85]	0.8439
Thyroid Medication	CLS	AUC	81 (2.5%)	0.930 [0.86–0.98]	0.956 [0.86–1.00]	0.4132

Supplementary Table 18 | **Classification Performance of Fitbit and SensorWM Features on Users Ages 40-59.** Classification performance comparing downstream prediction using true Fitbit features and world-model-generated Fitbit features for the age bucket 40–59. Values are AUROC with 95% confidence intervals. Statistical significance between AUROC estimates was assessed using DeLong's test.

Prediction Task	Type	Metric	Feature Source			
			N (Pos%)	True Fitbit Features	World Model Fitbit Features	DeLong p-val
Health Conditions						
Diabetes Condition	CLS	AUC	161 (9.9%)	0.731 [0.59–0.84]	0.736 [0.60–0.84]	0.8045
HOMA-IR Binary	CLS	AUC	69 (26.1%)	0.845 [0.74–0.94]	0.829 [0.71–0.93]	0.5330
Hyperlipidemia	CLS	AUC	161 (18.0%)	0.676 [0.56–0.79]	0.699 [0.59–0.80]	0.2625
Hypertension Binary	CLS	AUC	161 (18.0%)	0.766 [0.66–0.83]	0.783 [0.68–0.86]	0.3111
Respiratory Condition	CLS	AUC	161 (19.9%)	0.622 [0.53–0.72]	0.614 [0.52–0.72]	0.7114
Behavioral						
Smoking	CLS	AUC	160 (8.1%)	0.707 [0.57–0.83]	0.730 [0.59–0.86]	0.4850
Women's Health						
Regular Periods	CLS	AUC	94 (41.5%)	0.847 [0.77–0.92]	0.808 [0.70–0.90]	0.0242
Medication						
Cardiovascular Medication	CLS	AUC	177 (24.9%)	0.754 [0.67–0.83]	0.755 [0.67–0.84]	0.9123
Diabetes Medication	CLS	AUC	177 (10.7%)	0.644 [0.52–0.77]	0.631 [0.50–0.76]	0.3941
Mental Health Medication	CLS	AUC	177 (31.1%)	0.597 [0.50–0.70]	0.591 [0.51–0.70]	0.7604
No Medication	CLS	AUC	177 (41.8%)	0.713 [0.63–0.79]	0.682 [0.61–0.76]	0.0301
Thyroid Medication	CLS	AUC	177 (9.0%)	0.692 [0.53–0.83]	0.641 [0.50–0.78]	0.0382

Supplementary Table 19 | **Classification Performance of Fitbit and SensorWM Features on Users Ages 60-79.** Classification performance comparing downstream prediction using true Fitbit features and world-model-generated Fitbit features for the age bucket 60–79. Values are AUROC with 95% confidence intervals. Statistical significance between AUROC estimates was assessed using DeLong's test.

Prediction Task	Type	Metric	Feature Source			
			N (Pos%)	True Fitbit Features	World Model Fitbit Features	DeLong p-val
Health Conditions						
Diabetes Condition	CLS	AUC	68 (11.8%)	0.519 [0.40–0.67]	0.533 [0.41–0.68]	0.6658
HOMA-IR Binary	CLS	AUC	38 (26.3%)	0.686 [0.51–0.88]	0.618 [0.42–0.85]	0.0329
Hyperlipidemia	CLS	AUC	68 (36.8%)	0.476 [0.34–0.63]	0.463 [0.31–0.59]	0.6502
Hypertension Binary	CLS	AUC	68 (39.7%)	0.613 [0.49–0.75]	0.593 [0.45–0.73]	0.5674
Respiratory Condition	CLS	AUC	68 (20.6%)	0.669 [0.48–0.84]	0.679 [0.49–0.84]	0.5310
Behavioral						
Smoking	CLS	AUC	67 (1.5%)	0.864 [0.79–0.93]	0.848 [0.76–0.92]	—
Women's Health						
Regular Periods	CLS	AUC	37 (5.4%)	0.629 [0.46–0.80]	0.600 [0.46–0.80]	0.5953
Medication						
Cardiovascular Medication	CLS	AUC	71 (63.4%)	0.579 [0.42–0.71]	0.535 [0.39–0.67]	0.1262
Diabetes Medication	CLS	AUC	71 (14.1%)	0.708 [0.56–0.85]	0.707 [0.51–0.86]	0.9589
Mental Health Medication	CLS	AUC	71 (19.7%)	0.669 [0.50–0.84]	0.648 [0.46–0.83]	0.3585
No Medication	CLS	AUC	71 (23.9%)	0.517 [0.37–0.68]	0.486 [0.34–0.66]	0.2923
Thyroid Medication	CLS	AUC	71 (14.1%)	0.692 [0.52–0.87]	0.661 [0.51–0.86]	0.3490

Supplementary Table 20 | **Classification Performance of Fitbit and SensorWM Features on Users with a BMI <25** Classification performance comparing downstream prediction using true Fitbit features and world-model-generated Fitbit features for the BMI bucket < 25. Values are AUROC with 95% confidence intervals. Statistical significance between AUROC estimates was assessed using DeLong's test.

Prediction Task	Type	Metric	Feature Source			
			N (Pos%)	True Fitbit Features	World Model Fitbit Features	DeLong p-val
Health Conditions						
Diabetes Condition	CLS	AUC	105 (3.8%)	0.738 [0.60–0.91]	0.666 [0.54–0.83]	0.0057
HOMA-IR Binary	CLS	AUC	50 (10.0%)	0.800 [0.59–0.95]	0.809 [0.59–0.97]	0.9028
Hyperlipidemia	CLS	AUC	105 (10.5%)	0.617 [0.41–0.77]	0.620 [0.42–0.78]	0.8651
Hypertension Binary	CLS	AUC	105 (13.3%)	0.919 [0.86–0.97]	0.908 [0.85–0.97]	0.3072
Respiratory Condition	CLS	AUC	105 (9.5%)	0.544 [0.36–0.70]	0.575 [0.39–0.73]	0.2395
Behavioral						
Smoking	CLS	AUC	104 (9.6%)	0.807 [0.63–0.92]	0.760 [0.60–0.88]	0.2884
Women's Health						
Regular Periods	CLS	AUC	49 (40.8%)	0.855 [0.75–0.94]	0.848 [0.74–0.94]	0.6268
Medication						
Cardiovascular Medication	CLS	AUC	110 (24.5%)	0.816 [0.70–0.92]	0.791 [0.67–0.89]	0.0205
Diabetes Medication	CLS	AUC	110 (3.6%)	0.776 [0.55–1.00]	0.705 [0.45–0.99]	0.0375
Mental Health Medication	CLS	AUC	110 (17.3%)	0.665 [0.53–0.81]	0.636 [0.50–0.76]	0.2815
No Medication	CLS	AUC	110 (53.6%)	0.702 [0.62–0.79]	0.669 [0.58–0.78]	0.1063
Thyroid Medication	CLS	AUC	110 (3.6%)	0.825 [0.71–0.97]	0.842 [0.72–0.99]	0.2457

Supplementary Table 21 | **Classification Performance of Fitbit and SensorWM Features on Users with a BMI 25–30** Classification performance comparing downstream prediction using true Fitbit features and world-model-generated Fitbit features for the BMI bucket **25–30**. Values are AUROC with 95% confidence intervals. Statistical significance between AUROC estimates was assessed using DeLong's test.

Prediction Task	Type	Metric	Feature Source			
			N (Pos%)	True Fitbit Features	World Model Fitbit Features	DeLong p-val
Health Conditions						
Diabetes Condition	CLS	AUC	101 (6.9%)	0.713 [0.52–0.87]	0.679 [0.47–0.89]	0.3086
HOMA-IR Binary	CLS	AUC	53 (30.2%)	0.694 [0.52–0.83]	0.664 [0.51–0.81]	0.3352
Hyperlipidemia	CLS	AUC	101 (25.7%)	0.761 [0.67–0.85]	0.758 [0.68–0.86]	0.8727
Hypertension Binary	CLS	AUC	101 (17.8%)	0.744 [0.62–0.89]	0.723 [0.59–0.86]	0.4795
Respiratory Condition	CLS	AUC	101 (21.8%)	0.627 [0.48–0.77]	0.622 [0.46–0.77]	0.8209
Behavioral						
Smoking	CLS	AUC	99 (2.0%)	0.732 [0.51–0.93]	0.716 [0.50–0.90]	0.5847
Women's Health						
Regular Periods	CLS	AUC	47 (40.4%)	0.838 [0.71–0.94]	0.846 [0.74–0.95]	0.7195
Medication						
Cardiovascular Medication	CLS	AUC	108 (33.3%)	0.795 [0.69–0.90]	0.785 [0.69–0.88]	0.4623
Diabetes Medication	CLS	AUC	108 (8.3%)	0.707 [0.54–0.86]	0.681 [0.48–0.85]	0.2734
Mental Health Medication	CLS	AUC	108 (38.0%)	0.664 [0.58–0.78]	0.641 [0.55–0.76]	0.2887
No Medication	CLS	AUC	108 (35.2%)	0.788 [0.67–0.87]	0.768 [0.65–0.86]	0.2004
Thyroid Medication	CLS	AUC	108 (11.1%)	0.748 [0.59–0.88]	0.726 [0.55–0.87]	0.2418

Supplementary Table 22 | **Classification Performance of Fitbit and SensorWM Features on Users with a BMI >30** Classification performance comparing downstream prediction using true Fitbit features and world-model-generated Fitbit features for the BMI bucket > 30. Values are AUROC with 95% confidence intervals. Statistical significance between AUROC estimates was assessed using DeLong's test.

Prediction Task	Type	Metric	Feature Source			
			N (Pos%)	True Fitbit Features	World Model Fitbit Features	DeLong p-val
Health Conditions						
Diabetes Condition	CLS	AUC	99 (15.2%)	0.656 [0.52–0.80]	0.688 [0.54–0.82]	0.2623
HOMA-IR Binary	CLS	AUC	38 (44.7%)	0.711 [0.55–0.84]	0.675 [0.51–0.84]	0.4551
Hyperlipidemia	CLS	AUC	99 (23.2%)	0.662 [0.57–0.79]	0.668 [0.58–0.80]	0.7222
Hypertension Binary	CLS	AUC	99 (31.3%)	0.685 [0.57–0.78]	0.701 [0.60–0.80]	0.4107
Respiratory Condition	CLS	AUC	99 (25.3%)	0.562 [0.42–0.69]	0.542 [0.42–0.68]	0.5405
Behavioral						
Smoking	CLS	AUC	97 (10.3%)	0.786 [0.64–0.90]	0.746 [0.59–0.87]	0.4425
Women's Health						
Regular Periods	CLS	AUC	71 (38.0%)	0.850 [0.74–0.94]	0.826 [0.71–0.91]	0.1579
Medication						
Cardiovascular Medication	CLS	AUC	113 (34.5%)	0.731 [0.63–0.81]	0.731 [0.63–0.82]	0.9766
Diabetes Medication	CLS	AUC	113 (16.8%)	0.631 [0.49–0.78]	0.622 [0.48–0.77]	0.6933
Mental Health Medication	CLS	AUC	113 (32.7%)	0.542 [0.42–0.67]	0.544 [0.44–0.65]	0.9287
No Medication	CLS	AUC	113 (30.1%)	0.569 [0.44–0.70]	0.579 [0.44–0.71]	0.6592
Thyroid Medication	CLS	AUC	113 (10.6%)	0.702 [0.55–0.83]	0.634 [0.47–0.77]	0.0353

Supplementary Table 23 | **Classification Performance of Fitbit and SensorWM Features on Male Users** Classification performance comparing downstream prediction using true Fitbit features and world-model-generated Fitbit features for male users. Values are AUROC with 95% confidence intervals. Statistical significance between AUROC estimates was assessed using DeLong's test.

Prediction Task	Type	Metric	Feature Source			
			N (Pos%)	True Fitbit Features	World Model Fitbit Features	DeLong p-val
Health Conditions						
Diabetes Condition	CLS	AUC	131 (11.5%)	0.763 [0.64–0.86]	0.739 [0.61–0.83]	0.2654
HOMA-IR Binary	CLS	AUC	67 (29.9%)	0.815 [0.71–0.90]	0.797 [0.70–0.88]	0.4511
Hyperlipidemia	CLS	AUC	131 (23.7%)	0.664 [0.58–0.75]	0.675 [0.57–0.76]	0.3843
Hypertension Binary	CLS	AUC	131 (23.7%)	0.843 [0.75–0.91]	0.848 [0.76–0.92]	0.5870
Respiratory Condition	CLS	AUC	131 (10.7%)	0.747 [0.63–0.87]	0.725 [0.61–0.85]	0.1468
Behavioral						
Smoking	CLS	AUC	129 (5.4%)	0.905 [0.78–0.99]	0.854 [0.74–0.97]	0.2366
Medication						
Cardiovascular Medication	CLS	AUC	139 (33.1%)	0.875 [0.80–0.93]	0.871 [0.80–0.93]	0.5765
Diabetes Medication	CLS	AUC	139 (10.1%)	0.762 [0.65–0.84]	0.728 [0.60–0.82]	0.0597
Mental Health Medication	CLS	AUC	139 (21.6%)	0.617 [0.51–0.71]	0.594 [0.48–0.69]	0.3424
No Medication	CLS	AUC	139 (48.9%)	0.789 [0.70–0.86]	0.773 [0.69–0.84]	0.2678
Thyroid Medication	CLS	AUC	139 (3.6%)	0.687 [0.51–0.92]	0.649 [0.46–0.91]	0.3696

Supplementary Table 24 | **Classification Performance of Fitbit and SensorWM Features on Female Users** Classification performance comparing downstream prediction using true Fitbit features and world-model-generated Fitbit features for female users. Values are AUROC with 95% confidence intervals. Statistical significance between AUROC estimates was assessed using DeLong's test.

Prediction Task	Type	Metric	Feature Source			
			N (Pos%)	True Fitbit Features	World Model Fitbit Features	DeLong p-val
Health Conditions						
Diabetes Condition	CLS	AUC	174 (6.3%)	0.738 [0.63–0.87]	0.756 [0.64–0.87]	0.5171
HOMA-IR Binary	CLS	AUC	74 (24.3%)	0.752 [0.62–0.88]	0.730 [0.59–0.86]	0.4309
Hyperlipidemia	CLS	AUC	174 (16.7%)	0.720 [0.60–0.83]	0.714 [0.60–0.81]	0.6873
Hypertension Binary	CLS	AUC	174 (18.4%)	0.759 [0.64–0.85]	0.747 [0.64–0.82]	0.5723
Respiratory Condition	CLS	AUC	174 (24.7%)	0.554 [0.46–0.63]	0.562 [0.47–0.63]	0.6274
Behavioral						
Smoking	CLS	AUC	171 (8.8%)	0.704 [0.56–0.81]	0.685 [0.56–0.81]	0.6094
Women's Health						
Regular Periods	CLS	AUC	167 (39.5%)	0.852 [0.79–0.91]	0.834 [0.77–0.90]	0.0328
Medication						
Cardiovascular Medication	CLS	AUC	192 (29.2%)	0.716 [0.64–0.80]	0.697 [0.61–0.79]	0.1015
Diabetes Medication	CLS	AUC	192 (9.4%)	0.713 [0.61–0.81]	0.699 [0.58–0.80]	0.4083
Mental Health Medication	CLS	AUC	192 (34.9%)	0.613 [0.53–0.71]	0.592 [0.51–0.69]	0.3061
No Medication	CLS	AUC	192 (32.8%)	0.615 [0.54–0.68]	0.603 [0.53–0.68]	0.4308
Thyroid Medication	CLS	AUC	192 (12.0%)	0.699 [0.61–0.78]	0.652 [0.57–0.73]	0.0638

C.3. Expert Evaluation of Counterfactual Coaching

Verbatim Prompt 1: Generic Coaching Arm (Baseline)

Prompt for Personalized 12-Week Lifestyle Coaching

```
### ROLE & OBJECTIVE
You are an expert, certified Health and Wellness Coach grounded in
Self-Determination Theory (SDT) and Motivational Interviewing.  Your
objective is to write a personalized, supportive, and actionable 12-week
lifestyle coaching plan (500-700 words) for a participant based on their
demographic profile and 24-week wearable history.

### PARTICIPANT CONTEXT
* Demographics:  Age:  {age} years | Sex:  {sex} | BMI: {bmi} kg/m2
* Recent Measured Baseline (Daily Averages):
- Resting Heart Rate (RHR): {baseline_rhr} bpm
- Daily Steps:  {baseline_steps} steps/day
- Nightly Sleep Duration:  {baseline_sleep} hours/night
- Daily Cardio Workouts:  {baseline_cardio} mins/day
- Active Zone Minutes (AZM): {baseline_azm} mins/day
* 24-Week Habitual Distribution & Historical Peaks:
- Steps:  24-wk Average {hab_steps_mean}
(Range:  {hab_steps_min}-{hab_steps_max}, Std:  +/-{hab_steps_std}) |
Best Recorded Peak:  {peak_steps} steps/day
- Sleep:  24-wk Average {hab_sleep_mean}h
(Range:  {hab_sleep_min}-{hab_sleep_max}h, Std:  +/-{hab_sleep_std}h) |
Best Recorded Peak:  {peak_sleep} hours/night
- Cardio:  24-wk Average {hab_cardio_mean} mins
(Range:  {hab_cardio_min}-{hab_cardio_max} mins) | Best Recorded Peak:
{peak_cardio} mins/day

### COACHING & CLINICAL GUARDRAILS
1.  Tone:  Warm, empathetic, collaborative, and autonomy-supportive.
Never use guilt, fear, or shame.  Celebrate existing strengths before
suggesting improvements.
2.  Jargon Ban:  Do not use clinical acronyms, variable names, or
statistical terms (e.g., do not say "std", "min/max", or "BMI class").
Translate Resting Heart Rate (RHR) and Active Zone Minutes (AZM) into
clear, everyday concepts of heart efficiency and energizing movement.
3.  Safe & Realistic Calibration:  Anchor all behavioral goals within
the participant's current baseline and proven 24-week historical
peaks.  If BMI is elevated or baseline activity is sedentary, recommend
joint-friendly movement options (e.g., brisk walking, swimming, cycling).
4.  Length:  Keep the complete response between 500 and 700 words.

### REQUIRED 4-SECTION STRUCTURE
1.  YOUR HEALTH SNAPSHOT
- Warmly welcome the participant and synthesize their recent wearable
metrics (Steps, Sleep, RHR, and AZM) alongside their 24-week history
into a cohesive picture of where their body is thriving and where their
biggest opportunity lies.
- Reference their historical peaks as encouraging proof of what their
body has already achieved.
2.  YOUR 12-WEEK PROGRESSIVE PLAN
- Design a realistic, step-by-step 12-week roadmap divided into three
4-week phases:
```

Prompt for Personalized 12-Week Lifestyle Coaching (continued)

```
* Phase 1 (Weeks 1-4: Foundation): Gentle, low-friction adjustments to
build consistency around sleep and daily movement.
* Phase 2 (Weeks 5-8: Build): Gradual progression toward their
historical capacity, introducing moderate heart-pumping movement (AZM).
* Phase 3 (Weeks 9-12: Sustain): Consolidating habits into a
sustainable long-term lifestyle routine.
3. EVIDENCE-BASED HABIT FORMATION
- Provide concrete behavioral strategies tailored to this plan,
including Habit Stacking (anchoring new habits to existing daily cues),
Implementation Intentions ("If-Then" planning), Environmental Design, and
compassionate troubleshooting for busy or low-energy days.
4. SUMMARY & NEXT STEPS
- Conclude with an encouraging summary and a memorable one-sentence
takeaway that reinforces their agency and partnership.
```

Verbatim Prompt 2: Counterfactual Coaching Arm

Prompt for Counterfactual 12-Week Lifestyle Coaching

```
### ROLE & OBJECTIVE
You are an expert, certified Health and Wellness Coach grounded in
Self-Determination Theory (SDT) and Motivational Interviewing.  Your
objective is to write a personalized, supportive, and actionable
12-week lifestyle coaching plan (500-700 words) for a participant using
their 24-week wearable history, their Coachable Health Age, and their
personalized 12-week counterfactual health simulations.

### PARTICIPANT CONTEXT & COUNTERFACTUAL SIMULATIONS
* Demographics:  Age:  {age} years | Sex:  {sex} | BMI: {bmi} kg/m2
* Recent Measured Baseline (Daily Averages):
- Chronological Age:  {age} years
- Current Coachable Health Age:  {baseline_health_age} years
- Resting Heart Rate (RHR): {baseline_rhr} bpm
- Daily Steps:  {baseline_steps} steps/day
- Nightly Sleep Duration:  {baseline_sleep} hours/night
- Daily Cardio Workouts:  {baseline_cardio} mins/day
- Active Zone Minutes (AZM): {baseline_azm} mins/day
* 24-Week Habitual Distribution & Historical Peaks:
- Steps:  24-wk Average {hab_steps_mean}
(Range:  {hab_steps_min}-{hab_steps_max}, Std:  +/-{hab_steps_std}) |
Best Recorded Peak:  {peak_steps} steps/day
- Sleep:  24-wk Average {hab_sleep_mean}h
(Range:  {hab_sleep_min}-{hab_sleep_max}h, Std:  +/-{hab_sleep_std}h) |
Best Recorded Peak:  {peak_sleep} hours/night
- Cardio:  24-wk Average {hab_cardio_mean} mins
(Range:  {hab_cardio_min}-{hab_cardio_max} mins) | Best Recorded Peak:
{peak_cardio} mins/day
* 12-Week Status-Quo Projection (Future A - If Current Habits Continue
Unchanged):
- Projected Coachable Health Age:  {projected_ha_status_quo} years
- Projected Resting Heart Rate:  {projected_rhr_status_quo} bpm
- Projected Active Zone Minutes:  {projected_azm_status_quo} mins/day
* 12-Week Personalized Counterfactual Pathways (Future B - Simulated
Habit Options Bounded by Historical Peaks):
- Option 1 ({option_1_name}):  {option_1_habit_delta} -> Projected Health
Age:  {option_1_ha} years ({option_1_ha_delta} yrs), Projected RHR:
{option_1_rhr} bpm
- Option 2 ({option_2_name}):  {option_2_habit_delta} -> Projected Health
Age:  {option_2_ha} years ({option_2_ha_delta} yrs), Projected RHR:
{option_2_rhr} bpm
- Option 3 ({option_3_name}):  {option_3_habit_delta} -> Projected Health
Age:  {option_3_ha} years ({option_3_ha_delta} yrs), Projected RHR:
{option_3_rhr} bpm

### COACHING & CLINICAL GUARDRAILS
1.  Tone:  Warm, empathetic, collaborative, and autonomy-supportive.
Frame Coachable Health Age and Future A vs.  Future B as an empowering
compass-never use fear, guilt, or alarmism.
2.  Jargon Ban:  Never use the word "counterfactual", "simulation",
"model", "std", or "min/max".  Present Future A and Future B naturally
as "Where Your Current Routine Leads" versus "Your Personalized Pathways
Forward".
```

Prompt for Counterfactual 12-Week Lifestyle Coaching (continued)

```
3. Autonomy & Choice Architecture: Present the 2-3 pathways clearly so
the participant feels complete ownership over choosing the path (Sleep,
Movement, or Combined) that best fits their life right now.
4. Safe & Realistic Calibration: All pathways are bounded by the
participant's own 24-week historical peaks; reinforce that their
body has already demonstrated it can achieve these levels. Recommend
joint-friendly modalities when appropriate.
5. Length: Keep the complete response between 500 and 700 words.

### REQUIRED 4-SECTION STRUCTURE
1. YOUR HEALTH SNAPSHOT & HEALTH AGE
- Warmly welcome the participant and synthesize their wearable metrics
(Steps, Sleep, RHR, AZM) and Coachable Health Age into a unified compass.
- Gently contrast Future A (what happens over the next 12 weeks if
current habits stay the same) with Future B (how modifying specific daily
levers can lower their Health Age and Resting Heart Rate), referencing
their 24-week historical peaks as proof of capability.
2. YOUR PERSONALIZED PATHWAYS & 12-WEEK PROGRESSIVE PLAN
- Present the distinct, quantified Options (Option 1, Option 2, Option
3) with their exact 12-week Health Age and RHR payoffs, inviting the
participant to choose their preferred focus.
- Lay out a progressive 12-week roadmap divided into three 4-week phases:
* Phase 1 (Weeks 1-4: Foundation): Starting gently on their chosen
pathway.
* Phase 2 (Weeks 5-8: Build): Stepping up toward their target habit
change and heart-pumping movement (AZM).
* Phase 3 (Weeks 9-12: Sustain): Locking in the routine to realize
their 12-week Health Age and RHR improvements.
3. EVIDENCE-BASED HABIT FORMATION
- Provide concrete behavioral strategies (Habit Stacking, "If-Then"
Implementation Intentions, Self-Monitoring of RHR/energy, and
compassionate barrier planning) to support long-term adherence.
4. SUMMARY & NEXT STEPS
- Conclude with an encouraging summary and a memorable one-sentence
takeaway celebrating their control over their future health trajectory.
```

Supplementary Table 25 | **Baseline Characteristics of the Five WEAR-ME Coachable Personas.** Summary of demographic, physiological, and wearable-derived characteristics across the five personas, highlighting differences in baseline health profiles and activity patterns.

Persona ID & Clinical Archetype	**Selection Criteria (Steps, Sleep, BMI, ΔHA)**	**Full WEAR-ME Sub-Cohort (n, % of 2,780)**	**Sub-Cohort Means (Age, Steps, Sleep, BMI, RHR, ΔHA)**	**Evaluated Sample (n Valid Cases; PIDs)**	**Evaluated Sample Baseline (Mean ± SD)**	**Primary Counterfactual Coaching Test**
P1: Optimal Baseline *"The High-Vitality Longevity Pacer"*	Steps: ≥ 10k/day Sleep: 7.0–9.0 h BMI: 18.5–24.9 ΔHA: < −1.0 yr	n = 78 (2.81%)	Age: 53.0 yrs Steps: 13,775 Sleep: 7.42 h BMI: 22.9 kg/m^2 RHR: 60.1 bpm ΔHA: −4.99 yrs	n = 10 12611, 72404, 58662, 55814, 56084, 55957, 71670, 72398, 71845, 56638	Age: 57.6 ± 7.2 yrs Female: 40.0% (4/10) BMI: 23.2 ± 0.7 kg/m^2 Steps: 13, 111 ± 2, 455 Sleep: 7.22 ± 0.44 h RHR: 59.1 ± 3.5 bpm HA: 56.4 ± 7.2 yrs ΔHA: −1.22 ± 0.16 yrs	**Ceiling Effects & Maintenance:** Preventing overtraining and framing vitality preservation against age-related drift.
P2: Multi-Risk Challenger *"The Sedentary Fatigued Metabolic Seeker"*	Steps: < 5k/day Sleep: < 7.0 h BMI: ≥ 25.0 ΔHA: ≥ +1.0 yr	n = 277 (9.96%)	Age: 49.8 yrs Steps: 3,341 Sleep: 6.03 h BMI: 36.3 kg/m^2 RHR: 65.4 bpm ΔHA: +3.06 yrs	n = 10 186449, 56440, 13533, 70268, 70494, 56314, 57778, 58098, 72185, 56458	Age: 49.4 ± 4.6 yrs Female: 70.0% (7/10) BMI: 40.6 ± 5.9 kg/m^2 Steps: 3, 415 ± 1, 862 Sleep: 5.31 ± 1.38 h RHR: 68.8 ± 3.0 bpm HA: 57.4 ± 4.9 yrs ΔHA: +8.04 ± 1.09 yrs	**Multi-Lever Prioritization:** Avoiding cognitive overwhelm by offering low-friction entry pathways (Sleep vs. Steps) in high-risk patients.

Continued on next page

Supplementary Table 25 continued from previous page

Persona ID & Clinical Archetype	Selection Criteria (Steps, Sleep, BMI, ΔHA)	Full WEAR-ME Sub-Cohort (n, % of 2,780)	Sub-Cohort Means (Age, Steps, Sleep, BMI, RHR, ΔHA)	Evaluated Sample (n Valid Cases; PIDs)	Evaluated Sample Baseline (Mean ± SD)	Primary Counterfactual Coaching Test
P3: Recovery Bottleneck *"The Active Striver at Burnout Risk"*	Steps: ≥ 10k/day Sleep: < 7.0 h BMI: Healthy/OW ΔHA: Youthful/Sync	n = 354 (12.73%)	Age: 49.1 yrs Steps: 12,738 Sleep: 6.23 h BMI: 26.9 kg/m^2 RHR: 62.6 bpm ΔHA: −4.64 yrs	n = 10 70621, 71746, 71100, 58093, 70552, 13851, 55974, 70488, 71651, 58419	Age: 48.6 ± 2.5 yrs Female: 10.0% (1/10) BMI: 26.6 ± 1.3 kg/m^2 Steps: 12, 404 ± 3, 439 Sleep: 5.92 ± 1.54 h RHR: 64.7 ± 5.8 bpm HA: 49.2 ± 4.1 yrs ΔHA: +0.65 ± 2.84 yrs	**Lever Discrimination:** Recognizing saturated activity volume and isolating sleep extension and recovery rather than adding steps.
P4: Activity Bottleneck *"The Sedentary Age-Accelerated Individual"*	Steps: < 5k/day Sleep: 7.0–9.0 h BMI: OW/Healthy ΔHA: ≥ +1.0 yr	n = 227 (8.17%)	Age: 48.2 yrs Steps: 3,190 Sleep: 7.67 h BMI: 32.0 kg/m^2 RHR: 64.3 bpm ΔHA: +2.78 yrs	n = 10 12803, 13530, 13656, 56193, 72532, 72585, 72223, 71774, 57671, 58005	Age: 46.1 ± 4.9 yrs Female: 100.0% (10/10) BMI: 35.1 ± 4.7 kg/m^2 Steps: 2, 796 ± 788 Sleep: 6.95 ± 1.41 h RHR: 72.0 ± 7.7 bpm HA: 54.3 ± 4.6 yrs ΔHA: +8.17 ± 2.65 yrs	**Targeted Activity Escalation:** Isolating progressive step accumulation while validating and preserving healthy sleep habits.
P5: Balanced Maintainer *"The Balanced Health Seeker"*	Steps: 5k–10k/day Sleep: 7.0–9.0 h BMI: Healthy/OW ΔHA: −1.0 to +1.0	n = 143 (5.14%)	Age: 46.6 yrs Steps: 5,539 Sleep: 7.48 h BMI: 29.9 kg/m^2 RHR: 67.1 bpm ΔHA: −0.31 yrs	n = 9 valid[†] 55807, 56085, 56144, 58421, 70909, 57709, 57065, 58023, 57497	Age: 45.4 ± 5.0 yrs Female: 66.7% (6/9) BMI: 30.2 ± 3.5 kg/m^2 Steps: 5, 490 ± 2, 093 Sleep: 7.12 ± 1.38 h RHR: 69.7 ± 7.7 bpm HA: 50.1 ± 5.1 yrs ΔHA: +4.62 ± 3.66 yrs	**Fine-Tuning & Progression:** Nudging median compliant users from baseline maintenance into cardiovascular fitness gains.

Continued on next page

Supplementary Table 25 continued from previous page

Persona ID & Clinical Archetype	**Selection Criteria (Steps, Sleep, BMI, ΔHA)**	**Full WEAR-ME Sub-Cohort (*n*, % of 2,780)**	**Sub-Cohort Means (Age, Steps, Sleep, BMI, RHR, ΔHA)**	**Evaluated Sample (*n* Valid Cases; PIDs)**	**Evaluated Sample Baseline (Mean ± SD)**	**Primary Counterfactual Coaching Test**
Total / Overall	**5 Stratified Personas**	**1,079 / 2,780** (38.81%)	**Weighted Sub-Cohort**	$N = 49$ **valid** (50 sampled)	Age: 49.5 ± 6.6 yrs Female: 57.1% (28/49) BMI: 31.2 ± 7.2 kg/m^2 Steps: 7,483 ± 5,026 Sleep: 6.49 ± 1.46 h RHR: 66.8 ± 7.3 bpm HA: 53.6 ± 6.1 yrs ΔHA: +4.04 ± 4.52 yrs	**Comprehensive Evaluation Across the Risk Continuum**

† In Persona 5, Participant ID `57217` (Row 47) was excluded *a priori* because `Plan 1` and `Plan 2` contained identical text. Sub-cohort means reflect 24-week habitual means in the WEAR-ME persona registry, whereas evaluated sample means reflect the recent daily baseline snapshot presented to raters. OW, overweight.

Supplementary Table 26 | **Comprehensive Evaluation Results of the Lifestyle Coach.** Evaluation outcomes across five coaches and 49 valid cases, including head-to-head side-by-side (SxS) preference comparisons and score distributions across a 10-aspect evaluation rubric.

Evaluation Criterion / Statistic	**Counterfactual Coaching**	**Generic Coaching**
SxS: Best Overall Plan		
Scale	Forced-choice win	Forced-choice win
Full export (n = 245)	**71.4%** (175/245)	**28.6%** (70/245)
95% CI	65.5–76.7%	23.3–34.5%
Net difference	**+42.9 percentage points** favoring Counterfactual Coaching.	
Exact P-value	Two-sided exact binomial $P = 1.49 \times 10^{-11}$.	
Participant-level majority consensus	**43/49 cases (87.8%)**	6/49 cases (12.2%)
SxS: Highest Likelihood of Long-Term Adherence		
Scale	Forced-choice win	Forced-choice win
Full export (n = 245)	**71.0%** (174/245)	**29.0%** (71/245)
95% CI	65.0–76.3%	23.7–35.0%
Net difference	**+42.0 percentage points** favoring Counterfactual Coaching.	
Exact P-value	Two-sided exact binomial $P = 3.72 \times 10^{-11}$.	
Participant-level majority consensus	**44/49 cases (89.8%)**	5/49 cases (10.2%)
Contrasts Benefits vs. Status Quo		
Scale	Binary (Yes/No)	Binary (Yes/No)
Full export: Yes	**180/245 (73.5%)**	**14/245 (5.7%)**
95% CI	67.6–78.6%	3.4–9.4%
Full distribution: No	65/245 (26.5%)	231/245 (94.3%)
Net difference	**+67.8 percentage points** favoring Counterfactual Coaching.	
Discordant pairs	170	4
Exact McNemar P-value	$P = 3.15 \times 10^{-45}$.	
Compelling Reason to Change (No Fear/Shame)		
Scale	Binary (Yes/No)	Binary (Yes/No)
Full export: Yes	**185/245 (75.5%)**	**70/245 (28.6%)**
95% CI	69.8–80.5%	23.3–34.5%
Full distribution: No	60/245 (24.5%)	175/245 (71.4%)
Net difference	**+46.9 percentage points** favoring Counterfactual Coaching.	
Discordant pairs	121	6
Exact McNemar P-value	$P = 6.39 \times 10^{-29}$.	
Trajectory Control		
Scale	Binary (Yes/No)	Binary (Yes/No)
Full export: Yes	**238/245 (97.1%)**	**145/244 (59.4%)**‡
95% CI	94.2–98.6%	53.2–65.4%
Full distribution: No	7/245 (2.9%)	99/244 (40.6%)
Net difference	**+37.7 percentage points** favoring Counterfactual Coaching.	
Discordant pairs	97	5
Exact McNemar P-value	$P = 3.46 \times 10^{-23}$.	
Behavior-Change Strategies		
Scale	3-level ordinal	3-level ordinal
Top-box definition	Strongly Applied	Strongly Applied
Strongly Applied	**123/245 (50.2%)**	**50/245 (20.4%)**
95% CI for top-box	44.0–56.4%	15.8–25.9%
Superficial	121/245 (49.4%)	185/245 (75.5%)

Continued on next page

Supplementary Table 26 continued from previous page

Evaluation Criterion / Statistic	Counterfactual Coaching	Generic Coaching
Absent	1/245 (0.4%)	10/245 (4.1%)
Net top-box difference	**+29.8 percentage points** favoring Counterfactual Coaching.	
Discordant pairs	89	16
Exact McNemar *P*-value	$P = 1.87 \times 10^{-13}$.	
Sustainable Habit Building		
Scale	3-level ordinal	3-level ordinal
Top-box definition	Excellent	Excellent
Excellent	**128/245 (52.2%)**	**70/245 (28.6%)**
95% CI for top-box	46.0–58.4%	23.3–34.5%
Satisfactory	114/245 (46.5%)	167/245 (68.2%)
Poor	3/245 (1.2%)	8/245 (3.3%)
Net top-box difference	**+23.7 percentage points** favoring Counterfactual Coaching.	
Discordant pairs	75	17
Exact McNemar *P*-value	$P = 7.29 \times 10^{-10}$.	
Realistic & Non-Overwhelming Routine Fit		
Scale	Binary (Yes/No)	Binary (Yes/No)
Full export: Yes	**242/245 (98.8%)**	**203/245 (82.9%)**
95% CI	96.5–99.6%	77.6–87.1%
Full distribution: No	3/245 (1.2%)	42/245 (17.1%)
Net difference	**+15.9 percentage points** favoring Counterfactual Coaching.	
Discordant pairs	41	2
Exact McNemar *P*-value	$P = 2.15 \times 10^{-10}$.	
Supportive & Motivating Tone		
Scale	Binary (Yes/No)	Binary (Yes/No)
Full export: Yes	**241/245 (98.4%)**	**204/245 (83.3%)**
95% CI	95.9–99.4%	78.1–87.4%
Full distribution: No	4/245 (1.6%)	41/245 (16.7%)
Net difference	**+15.1 percentage points** favoring Counterfactual Coaching.	
Discordant pairs	41	4
Exact McNemar *P*-value	$P = 9.33 \times 10^{-9}$.	
Baseline & Wearable Personalization		
Scale	3-level ordinal	3-level ordinal
Top-box definition	Highly Tailored	Highly Tailored
Highly Tailored	**133/245 (54.3%)**	**127/245 (51.8%)**
95% CI for top-box	48.0–60.4%	45.6–58.0%
Partially Tailored	106/245 (43.3%)	108/245 (44.1%)
Generic	6/245 (2.4%)	10/245 (4.1%)[§]
Net top-box difference	**+2.4 percentage points** favoring Counterfactual Coaching.	
Discordant pairs	42	36
Exact McNemar *P*-value	$P = 0.572$; not statistically significant.	
Signal Synthesis (Unified Compass)		
Scale	3-level ordinal	3-level ordinal
Top-box definition	Unified Compass	Unified Compass
Unified Compass	**120/245 (49.0%)**	**119/245 (48.6%)**
95% CI for top-box	42.8–55.2%	42.4–54.8%
Descriptive Summary	106/245 (43.3%)	73/245 (29.8%)
Fragmented	19/245 (7.8%)[§]	48/245 (19.6%)
Net top-box difference	**+0.4 percentage points**.	
Discordant pairs	47	46

Continued on next page

Supplementary Table 26 continued from previous page

Evaluation Criterion / Statistic	**Counterfactual Coaching**	**Generic Coaching**
Exact McNemar *P*-value	$P = 1.000$; not statistically significant.	
Easy to Understand		
Scale	Binary (Yes/No)	Binary (Yes/No)
Full export: Yes	**208/245 (84.9%)**	**220/245 (89.8%)**
95% CI	79.9–88.8%	85.4–93.0%
Full distribution: No	37/245 (15.1%)	25/245 (10.2%)
Net difference	−4.9 **percentage points** for Counterfactual relative to Generic Coaching.	
Discordant pairs	18	30
Exact McNemar *P*-value	$P = 0.111$; not statistically significant.	

‡ One Generic `trajectory_control` rating was left blank by a coach, resulting in $n = 244$ for this item. In 5/245 Generic `wearable_synthesis` cells, a text comment was entered in place of a categorical rating.

§ Although top-box *Unified Compass* rates were similar (49.0% vs. 48.6%), Counterfactual Coaching reduced bottom-box *Fragmented* synthesis from 19.6% to 7.8%. Counterfactual Coaching also reduced bottom-box *Generic* personalization from 4.1% to 2.4%.

CF, Counterfactual Coaching; Gen, Generic Coaching; SxS, side-by-side comparison; CI, confidence interval. Unless otherwise specified, values are reported as Counterfactual versus Generic. Paired rubric comparisons were assessed using exact McNemar tests.

Supplementary Table 27 | **Comprehensive Evaluation Results of the Clinician Assessment.** Evaluation outcomes across five clinicians and 49 valid cases, comprising 245 paired evaluations across six dimensions assessing safety, accuracy, plausibility, and personalization.

Evaluation Criterion / Statistic	**Generic Coaching**	**Counterfactual Coaching**
Safe Intensity Limits (Primary Safety Gate)		
Scale	Binary (Yes/No)	Binary (Yes/No)
Fig. 1 & 5 bar values	**97.6% Yes** (panel 1e) **98% Yes** (panel 5b)	**99.2% Yes** (panel 1e) **99% Yes** (panel 5b)
Full export: Yes	**239/245 (97.6%)**	**243/245 (99.2%)**
95% CI	94.8–98.9%	97.1–99.8%
Full distribution: No	6/245 (2.4%)	2/245 (0.8%)
Net difference (CF−Gen)	**+1.6 percentage points**.	
Discordant pairs (b vs. c)	6	2
Exact McNemar P-value	$P = 0.289$; not statistically significant.	
Physiological Claims Correct		
Scale	Binary (Yes/No)	Binary (Yes/No)
Fig. 5 bar values	**98% Yes**	**97% Yes**
Full export: Yes	**239/245 (97.6%)**	**238/245 (97.1%)**
95% CI	94.8–98.9%	94.2–98.6%
Full distribution: No	6/245 (2.4%)	7/245 (2.9%)
Net difference (CF−Gen)	−0.4 **percentage points**.	
Discordant pairs (b vs. c)	2	3
Exact McNemar P-value	$P = 1.000$; not statistically significant.	
Trajectory Control		
Scale	Binary (Yes/No)	Binary (Yes/No)
Fig. 5 bar values	**93% Yes**	**97% Yes**
Full export: Yes	**228/245 (93.1%)**	**237/245 (96.7%)**
95% CI	89.2–95.6%	93.7–98.3%
Full distribution: No	17/245 (6.9%)	8/245 (3.3%)
Net difference (CF−Gen)	**+3.7 percentage points**.	
Discordant pairs (b vs. c)	16	7
Exact McNemar P-value	$P = 0.093$; not statistically significant.	
12-Week Biological Plausibility		
Scale	Binary (Yes/No)	Binary (Yes/No)
Fig. 5 bar values	**82% Yes**	**86% Yes**
Full export: Yes	**200/245 (81.6%)**	**211/245 (86.1%)**
95% CI	76.3–86.0%	81.2–89.9%
Full distribution: No	45/245 (18.4%)	34/245 (13.9%)
Net difference (CF−Gen)	**+4.5 percentage points**.	
Discordant pairs (b vs. c)	18	7
Exact McNemar P-value	$P = 0.043$.	
Contrasts Benefits vs. Status Quo		
Scale	Binary (Yes/No)	Binary (Yes/No)
Fig. 5 bar values	**8% Yes**	**92% Yes**
Full export: Yes	**19/245 (7.8%)**	**226/245 (92.2%)**
95% CI	5.0–11.8%	88.2–95.0%
Full distribution: No	226/245 (92.2%)	19/245 (7.8%)
Net difference (CF−Gen)	**+84.5 percentage points**.	
Discordant pairs (b vs. c)	213	6
Exact McNemar P-value	$P = 3.49 \times 10^{-55}$.	

Continued on next page

Supplementary Table 27 continued from previous page

Evaluation Criterion / Statistic	Generic Coaching	Counterfactual Coaching
Baseline & Wearable Personalization		
Scale	3-level ordinal	3-level ordinal
Top-box definition	Highly Tailored	Highly Tailored
Fig. 5 bar values	**52%**	**72%**
Highly Tailored	**128/245 (52.2%)**	**177/245 (72.2%)**
95% CI for top-box	46.0–58.4%	66.3–77.5%
Partially Tailored	109/245 (44.5%)	50/245 (20.4%)
Generic	8/245 (3.3%)	18/245 (7.3%)
Net top-box difference (CF−Gen)	**+20.0 percentage points**.	
Discordant pairs (*b* vs. *c*)	82	33
Exact McNemar *P*-value	$P = 5.57 \times 10^{-6}$.	

Gen, Generic Coaching; CF, Counterfactual Coaching; CI, confidence interval. Values are reported for $n = 245$ paired clinician evaluations unless otherwise stated.

Paired comparisons were assessed using two-sided exact McNemar tests. Exact *P*-values are reported directly; $P < 0.05$ was considered statistically significant.

Supplementary Table 28 | **Cross-Disciplinary Concordance and Divergence in Evaluation Outcomes.** Agreement and disagreement between lifestyle coaches and clinicians across shared evaluation dimensions, based on 490 pooled paired evaluations.

Shared Dimension / Statistic	**Lifestyle Coaches (n = 245)**	**Clinicians (n = 245)**
Contrasts Benefits vs. Status Quo		
Scale	Binary (% Yes)	Binary (% Yes)
Counterfactual	**73.5%**	**92.2%**
Generic	5.7%	7.8%
Net difference (CF−Gen)	**+67.8 percentage points**	**+84.5 percentage points**
Exact P-value	3.15×10^{-45}	3.49×10^{-55}
Pooled Counterfactual	**82.9% (406/490)**	
Pooled Generic	**6.7% (33/490)**	
Pooled difference	**+76.1 percentage points**; $P = 2.20 \times 10^{-99}$.	
Cross-disciplinary interpretation	**Strong Concordance:** Both expert groups showed the same qualitative pattern. Generic plans rarely contrasted the projected status quo with the benefits of behavioral change, whereas Counterfactual plans consistently provided this comparison.	
Trajectory Control		
Scale	Binary (% Yes)	Binary (% Yes)
Counterfactual	**97.1%**	**96.7%**
Generic	59.4%	93.1%
Net difference (CF−Gen)	**+37.7 percentage points**	**+3.7 percentage points**
Exact P-value	3.46×10^{-23}	0.093
Pooled Counterfactual	**96.9% (475/490)**	
Pooled Generic	**76.3% (373/489)**	
Pooled difference	**+20.7 percentage points**; $P = 9.25 \times 10^{-22}$.	
Cross-disciplinary interpretation	**Behavioral Stringency:** Both groups rated Counterfactual plans near ceiling for trajectory control (~ 97%). Lifestyle coaches, however, rated Generic plans substantially lower (59.4%), consistent with a stricter emphasis on explicit choice architecture and multiple behavioral pathways. Clinicians rated Generic plans highly (93.1%), indicating that a gradual, modifiable 12-week plan was often sufficient to satisfy their interpretation of trajectory control.	
Baseline & Wearable Personalization		
Scale	3-level ordinal (% Highly Tailored)	3-level ordinal (% Highly Tailored)
Counterfactual	**54.3%**	**72.2%**
Generic	51.8%	52.2%
Net difference (CF−Gen)	**+2.4 percentage points**	**+20.0 percentage points**
Exact P-value	0.572	5.57×10^{-6}
Pooled Counterfactual	**63.3% (310/490)**	
Pooled Generic	**52.0% (255/490)**	
Pooled difference	**+11.2 percentage points**; $P = 9.18 \times 10^{-5}$.	
Cross-disciplinary interpretation	**Clinical vs. Psychosocial Lens:** Clinicians showed a substantially larger personalization difference between Counterfactual and Generic plans, consistent with greater emphasis on physiological tailoring to Health Age and resting-heart-rate patterns. Lifestyle coaches rated the two conditions more similarly, reflecting a broader personalization standard that also considers psychosocial context such as life schedules, barriers, caregiving responsibilities, and access to exercise resources.	

Continued on next page

Supplementary Table 28 continued from previous page

Shared Dimension / Statistic	**Lifestyle Coaches (*n* = 245)**	**Clinicians (*n* = 245)**

CF, Counterfactual Coaching; Gen, Generic Coaching. Pooled values combine the 245 lifestyle-coach and 245 clinician evaluations. For Trajectory Control, the Generic pooled denominator is $n = 489$ because one lifestyle-coach Generic rating was missing.

The pooled statistics summarize the combined expert evaluations, whereas the discipline-specific columns show the corresponding results separately for lifestyle coaches and clinicians.

Supplementary Table 29 | **Stratified Evaluation Performance Across the Five WEAR-ME Coachable Personas.** Comprehensive breakdown of evaluation outcomes stratified by coachable persona, highlighting variations in performance across distinct user profiles and corresponding to the results presented in Fig. 5c.

Dimension / Endpoint	Coachable Persona	Result
Coach SxS: Best Overall Plan	*(Counterfactual vs. Generic; exact binomial P)*	
	P1: Optimal Baseline ≥ 10k steps/day; 7–9 h sleep $n = 50$	**72.0% vs. 28.0%** 36/50 $P = 0.0026$
	P2: Multi-Risk Challenger < 5k steps/day; < 7 h sleep; high BMI $n = 50$	**80.0% vs. 20.0%** 40/50 $P = 2.39 \times 10^{-5}$
	P3: Recovery Bottleneck ≥ 10k steps/day; < 7 h sleep $n = 50$	**60.0% vs. 40.0%** 30/50 $P = 0.203$
	P4: Activity Bottleneck < 5k steps/day; 7–9 h sleep $n = 50$	**74.0% vs. 26.0%** 37/50 $P = 0.0009$
	P5: Balanced Maintainer 5–10k steps/day; 7–9 h sleep $n = 45$; 9 cases	**71.1% vs. 28.9%** 32/45 $P = 0.0066$
Coach SxS: Long-Term Adherence	*(Counterfactual vs. Generic; exact binomial P)*	
	P1: Optimal Baseline	**70.0% vs. 30.0%** 35/50 $P = 0.0066$
	P2: Multi-Risk Challenger	**82.0% vs. 18.0%** 41/50 $P = 5.61 \times 10^{-6}$
	P3: Recovery Bottleneck	**58.0% vs. 42.0%** 29/50 $P = 0.322$
	P4: Activity Bottleneck	**74.0% vs. 26.0%** 37/50 $P = 0.0009$
	P5: Balanced Maintainer	**71.1% vs. 28.9%** 32/45 $P = 0.0066$
Trajectory Control (Lifestyle Coach)	Δ = *Counterfactual − Generic*	
	P1: Optimal Baseline	**+34.7 percentage points** 100% vs. 65%
	P2: Multi-Risk Challenger	**+48.0 percentage points** 100% vs. 52%
	P3: Recovery Bottleneck	**+24.0 percentage points** 96% vs. 72%
	P4: Activity Bottleneck	**+54.0 percentage points** 100% vs. 46%
	P5: Balanced Maintainer	**+26.7 percentage points** 89% vs. 62%
Behavior-Change Strategies* (Lifestyle Coach)	Δ = *Counterfactual − Generic*	

Continued on next page

Supplementary Table 29 continued from previous page

Dimension / Endpoint	Coachable Persona	Result
	P1: Optimal Baseline	**+24.0 percentage points** 58% vs. 34%
	P2: Multi-Risk Challenger	**+32.0 percentage points** 48% vs. 16%
	P3: Recovery Bottleneck	**+18.0 percentage points** 44% vs. 26%
	P4: Activity Bottleneck	**+34.0 percentage points** 44% vs. 10%
	P5: Balanced Maintainer	**+42.2 percentage points** 58% vs. 16%
Sustainable Habit Building* **(Lifestyle Coach)**	Δ = *Counterfactual − Generic*	
	P1: Optimal Baseline	**+28.0 percentage points** 64% vs. 36%
	P2: Multi-Risk Challenger	**+26.0 percentage points** 50% vs. 24%
	P3: Recovery Bottleneck	**+12.0 percentage points** 48% vs. 36%
	P4: Activity Bottleneck	**+26.0 percentage points** 50% vs. 24%
	P5: Balanced Maintainer	**+26.7 percentage points** 49% vs. 22%
Realistic & Manageable Routine Fit (Lifestyle Coach)	Δ = *Counterfactual − Generic*	
	P1: Optimal Baseline	**+16.0 percentage points** 98% vs. 82%
	P2: Multi-Risk Challenger	**+26.0 percentage points** 100% vs. 74%
	P3: Recovery Bottleneck	**+4.0 percentage points** 98% vs. 94%
	P4: Activity Bottleneck	**+14.0 percentage points** 100% vs. 86%
	P5: Balanced Maintainer	**+20.0 percentage points** 98% vs. 78%
Baseline Personalization* **(Lifestyle Coach)**	Δ = *Counterfactual − Generic*	
	P1: Optimal Baseline	**+10.0 percentage points** 70% vs. 60%
	P2: Multi-Risk Challenger	**+4.0 percentage points** 48% vs. 44%
	P3: Recovery Bottleneck	−14.0 **percentage points** 44% vs. 58%
	P4: Activity Bottleneck	**+8.0 percentage points** 56% vs. 48%
	P5: Balanced Maintainer	**+4.4 percentage points** 53% vs. 49%
Contrasts Benefits vs. Status Quo (Lifestyle Coach)	Δ = *Counterfactual − Generic*	
	P1: Optimal Baseline	**+72.0 percentage points** 78% vs. 6%
	P2: Multi-Risk Challenger	**+68.0 percentage points** 72% vs. 4%
	P3: Recovery Bottleneck	**+66.0 percentage points** 72% vs. 6%

Continued on next page

Supplementary Table 29 continued from previous page

Dimension / Endpoint	Coachable Persona	Result
	P4: Activity Bottleneck	**+72.0 percentage points** 74% vs. 2%
	P5: Balanced Maintainer	**+60.0 percentage points** 71% vs. 11%
Compelling Reason to Change (Lifestyle Coach)	Δ = *Counterfactual − Generic*	
	P1: Optimal Baseline	**+42.0 percentage points** 78% vs. 36%
	P2: Multi-Risk Challenger	**+54.0 percentage points** 72% vs. 18%
	P3: Recovery Bottleneck	**+44.0 percentage points** 76% vs. 32%
	P4: Activity Bottleneck	**+54.0 percentage points** 78% vs. 24%
	P5: Balanced Maintainer	**+40.0 percentage points** 73% vs. 33%
Supportive & Motivating Tone (Lifestyle Coach)	Δ = *Counterfactual − Generic*	
	P1: Optimal Baseline	**+16.0 percentage points** 98% vs. 82%
	P2: Multi-Risk Challenger	**+16.0 percentage points** 98% vs. 82%
	P3: Recovery Bottleneck	**+14.0 percentage points** 100% vs. 86%
	P4: Activity Bottleneck	**+18.0 percentage points** 100% vs. 82%
	P5: Balanced Maintainer	**+11.1 percentage points** 96% vs. 84%
Signal Synthesis* (Lifestyle Coach)	Δ = *Counterfactual − Generic*	
	P1: Optimal Baseline	**+10.0 percentage points** 56% vs. 46%
	P2: Multi-Risk Challenger	−2.0 **percentage points** 40% vs. 42%
	P3: Recovery Bottleneck	**+4.0 percentage points** 54% vs. 50%
	P4: Activity Bottleneck	−2.0 **percentage points** 46% vs. 48%
	P5: Balanced Maintainer	−8.9 **percentage points** 49% vs. 58%
Easy to Understand (Lifestyle Coach)	Δ = *Counterfactual − Generic*	
	P1: Optimal Baseline	**0.0 percentage points** 94% vs. 94%
	P2: Multi-Risk Challenger	−8.0 **percentage points** 78% vs. 86%
	P3: Recovery Bottleneck	−10.0 **percentage points** 84% vs. 94%
	P4: Activity Bottleneck	−4.0 **percentage points** 78% vs. 82%
	P5: Balanced Maintainer	−2.2 **percentage points** 91% vs. 93%
Baseline Personalization* (Clinician)	Δ = *Counterfactual − Generic*	
	P1: Optimal Baseline	**+16.0 percentage points** 76% vs. 60%

Continued on next page

Supplementary Table 29 continued from previous page

Dimension / Endpoint	Coachable Persona	Result
	P2: Multi-Risk Challenger	**+2.0 percentage points** 62% vs. 60%
	P3: Recovery Bottleneck	**+18.0 percentage points** 70% vs. 52%
	P4: Activity Bottleneck	**+38.0 percentage points** 76% vs. 38%
	P5: Balanced Maintainer	**+26.7 percentage points** 78% vs. 51%
12-Week Biological Plausibility (Clinician)	Δ = *Counterfactual − Generic*	
	P1: Optimal Baseline	**+18.0 percentage points** 98% vs. 80%
	P2: Multi-Risk Challenger	−6.0 **percentage points** 76% vs. 82%
	P3: Recovery Bottleneck	**+6.0 percentage points** 92% vs. 86%
	P4: Activity Bottleneck	**+2.0 percentage points** 82% vs. 80%
	P5: Balanced Maintainer	**+2.2 percentage points** 82% vs. 80%
Safe Intensity Limits (Clinician)	Δ = *Counterfactual − Generic*	
	P1: Optimal Baseline	**+2.0 percentage points** 100% vs. 98%
	P2: Multi-Risk Challenger	**+2.0 percentage points** 100% vs. 98%
	P3: Recovery Bottleneck	**+2.0 percentage points** 98% vs. 96%
	P4: Activity Bottleneck	**+2.0 percentage points** 100% vs. 98%
	P5: Balanced Maintainer	**0.0 percentage points** 98% vs. 98%
Contrasts Benefits vs. Status Quo (Clinician)	Δ = *Counterfactual − Generic*	
	P1: Optimal Baseline	**+96.0 percentage points** 100% vs. 4%
	P2: Multi-Risk Challenger	**+84.0 percentage points** 90% vs. 6%
	P3: Recovery Bottleneck	**+84.0 percentage points** 90% vs. 6%
	P4: Activity Bottleneck	**+88.0 percentage points** 96% vs. 8%
	P5: Balanced Maintainer	**+68.9 percentage points** 84% vs. 16%

* Rubric names marked with an asterisk denote 3-level ordinal measures for which the reported percentage corresponds to top-box achievement.
For the side-by-side preference outcomes, the original analysis used $^{**}P < 0.01$ and $^{***}P < 0.001$ as significance thresholds. Exact *P*-values are reported directly here.
CF, Counterfactual Coaching; Gen, Generic Coaching. Unless otherwise indicated, percentages within the Result column are reported as Counterfactual versus Generic.

Supplementary Table 30 | **Individual Expert Rater Performance and Qualitative Feedback.** Detailed evaluation outcomes for individual lifestyle coaches and clinicians, accompanied by representative verbatim feedback highlighting expert perspectives on model performance, strengths, and limitations.

Expert Track & Rater ID	Counterfactual Win Rate / Key Rubric Highlights	Representative Verbatim Qualitative Feedback (Unblinded Post-Analysis)
Lifestyle Coach 1 Valid cases: $n = 49$	**Best Overall:** 89.8% (44/49) **Adherence:** 98.0% (48/49) **Contrasts Benefits:** 98% CF vs. 2% Gen	*"Although both plans are helpful, supportive, and use data and behavior change techniques, [the Counterfactual plan] provides much more detail and dives deeper into personalizing the plan and providing options for the user to choose, giving the user the autonomy to select the best option for them which will be stronger in long-term success."* (PIDs 186449, 55807)
Lifestyle Coach 2 Valid cases: $n = 49$	**Best Overall:** 53.1% (26/49) **Adherence:** 44.9% (22/49) **Trajectory Control:** 92% CF vs. 67% Gen	*"[The Counterfactual plan] is by far more effective with the greatest likelihood to impact both short-term and long-term behavior. While both positively reinforce behavior at the outset, [the Generic plan] fails to contrast short- and long-term impacts nor does it explain the science-backed 'why'. In P3 (PID 70621), [Counterfactual] provides the best opportunity for user agency, whereas [Generic] misses sleep as the key factor for someone already at 10,000+ steps/day."*
Lifestyle Coach 3 Valid cases: $n = 49$	**Best Overall:** 98.0% (48/49) **Adherence:** 98.0% (48/49) **Compelling Reason:** 98% CF vs. 4% Gen	*"I loved how thorough the coach was with detailing what would happen to the user by choosing the plan and following it to their long-term benefits... [The Counterfactual plan] would give me the motivation to start the plan I want to choose and know what benefits could come from doing so. [The Generic plan] does not do that at all."* **Constructive critique on PID 186449:** *"Needs to clear up the phrasing of adding +1,000 steps per day so the user does not confuse a daily target with adding another 1,000 every single day."*
Lifestyle Coach 4 Valid cases: $n = 49$	**Best Overall:** 57.1% (28/49) **Adherence:** 53.1% (26/49) **Behavior Strategies:** 37% CF vs. 39% Gen	*"In [the Generic plan], there is not an active contrast here for the user, and there could have been more tangible points to connect the dots on where their health is heading if they choose to adopt this plan. [The Counterfactual plan] connects the metrics directly to future outcomes and gives the user tangible options."*

Continued on next page

Supplementary Table 30 continued from previous page

Expert Track & Rater ID	Counterfactual Win Rate / Key Rubric Highlights	Representative Verbatim Qualitative Feedback (Unblinded Post-Analysis)
Lifestyle Coach 5 Valid cases: $n = 49$	**Best Overall:** 59.2% (29/49) **Adherence:** 61.2% (30/49) **Trajectory Control:** 100% CF vs. 98% Gen	*"[The Counterfactual plan] is more flexible with multiple pathways, while [the Generic plan] is more rigid and increasingly takes more time for the user to complete, which can be a barrier to success. In [the Generic plan], each metric is touched upon separately with nothing that addresses the bigger 'why' for making these changes."*
Clinician 1 Valid cases: $n = 49$	**Safety Gate:** 100% CF vs. 94% Gen **Contrasts Benefits:** 98% CF vs. 2% Gen **Personalization:** 88% CF vs. 82% Gen	*"[The Counterfactual plan] is much more thorough in explaining the physiological consequences of staying the same versus making changes, and gives options of paths to follow that are flexible and the user's choice."*
Clinician 2 Valid cases: $n = 49$	**Safety Gate:** 100% CF vs. 100% Gen **Contrasts Benefits:** 98% CF vs. 29% Gen **Personalization:** 73% CF vs. 65% Gen	*"Overall [the Generic plan] is more vague; there are no suggestions of a path to choose and it does not give the user the numbers to motivate what the new habit will do for their health age. [The Counterfactual plan] clearly notes Future A and Future B and gives Options 1–3 as far as areas to choose focus."*
Clinician 3 Valid cases: $n = 49$	**Safety Gate:** 96% CF vs. 96% Gen **Bio. Plausibility:** 88% CF vs. 90% Gen **Personalization:** 73% CF vs. 14% Gen	*"[The Generic plan] lacks contrast between active lifestyle adoption and current level of engagement. [The Counterfactual plan] understands where the patient starts, proposes gentle increments bounded by their history, and makes accurate claims across sleep, exercise, and steps; with perseverance and commitment, the improved metrics are achievable."*
Clinician 4 Valid cases: $n = 49$	**Safety Gate:** 100% CF vs. 100% Gen **Bio. Plausibility:** 100% CF vs. 100% Gen **Contrasts Benefits:** 98% CF vs. 2% Gen	*"High clinical safety and physiological accuracy across both arms, with [the Counterfactual plan] providing superior quantification of expected RHR and Health Age improvements tailored to baseline bottlenecks."*
Clinician 5 Valid cases: $n = 49$	**Safety Gate:** 100% CF vs. 98% Gen **Contrasts Benefits:** 94% CF vs. 4% Gen **Bio. Plausibility:** 43% CF vs. 20% Gen	**Constructive clinical plausibility critique on Persona 2, PIDs `186449` and `56440`:** *"While the counterfactual options are motivating, dropping Health Age by 4–5 years in 12 weeks by only adding 30 minutes of sleep or modest steps is unlikely to happen that quickly in a severely deconditioned patient."*

CF, Counterfactual Coaching; Gen, Generic Coaching; SxS, side-by-side comparison; RHR, resting heart rate. Win rates and rubric highlights are reported at the individual expert-rater level.

Continued on next page

Supplementary Table 30 continued from previous page

Expert Track & Rater ID	Counterfactual Win Rate / Key Rubric Highlights	Representative Verbatim Qualitative Feedback (Unblinded Post-Analysis)
Qualitative comments are reproduced verbatim from the unblinded post-analysis feedback provided by lifestyle coaches and clinicians. Participant identifiers are shown only where they were included in the original qualitative feedback.		