MULTI-AGENT FRAMEWORK FOR DEVELOPING AND EVALUATING BIOMETRIC-BASED MULTIMODAL SYNTHETIC PERSONAS: A CASE STUDY ON PREGNANCY AND POSTPARTUM USERS

Poojita Garg[1,2,*], **Conor Heneghan**[1], **Katherine Heller**[1], **Taedong Yun**[1], **Maja Matarić**[1], **Mercy Nyamewaa Asiedu**[1]

[1]Google [2]University of Washington

ABSTRACT

Despite the rise of LLM-based health agents, robust and privacy-preserving evaluation frameworks remain scarce. We present a framework using a Transformer-based TimeGAN to generate synthetic biometric time-series from 3,497 real participant data. These time series are integrated with LLM-generated narrative vignettes to create biometrically-aligned multimodal personas for realistic user-agent simulations. Employing an LLM-as-a-Judge methodology, we evaluate agent performance across passive and proactive modes. Our results demonstrate that this framework effectively identifies critical gaps in current LLM capabilities regarding factual correctness, data usage, and clinical plausibility. Using pregnancy and postpartum period as demonstrative examples, our methodology provides a scalable tool for optimizing and evaluating personal health agents across diverse biometric and clinical applications.

Track: Research

1 INTRODUCTION

Large Language Model (LLM)-based personal health agents are witnessing rapid adoption, enabling users to interact with wearable biometric data for insights into sleep, activity, and heart rate. Recent research has focused on enhancing these integrations through query rewriting (Ren et al., 2025), physiological empathy (Dongre, 2024), and direct sensor-based health predictions (Kim et al., 2024).

Despite these advances, robust pipelines for evaluating multimodal (text and biometric time-series) agentic frameworks remain scarce. Privacy concerns regarding sensitive biomarkers limit the availability of high-quality open-source datasets, while accompanying use-case vignettes for diverse health conditions are often lacking. Current synthetic user research, such as SYNTHIA (Rahimzadeh et al., 2025), OpenCharacter (Wang et al., 2025), and Yun et al. (2025) provides strong narrative foundations but remains primarily monomodal (text-only). These frameworks lack the continuous biometric time-series data necessary to simulate the dynamic physiological shifts characteristic of real-world wearable users. Furthermore, existing evaluation benchmarks like PersonaGym (Samuel et al., 2024) and PersonaMem (Jiang et al., 2025) often prioritize general persona adherence over domain-specific factual correctness. Such approaches may overestimate an agent's practical ability to independently retrieve and apply complex biometric facts within free-form conversation.

To bridge these gaps, we present a multi-agent evaluation framework that generates multimodal synthetic personas by fusing rich textual vignettes with synthetic biometric time-series. We focus on the pregnancy and postpartum population, a cohort defined by distinct physiological changes and highly sensitive data needs. Leveraging an IRB-approved dataset from 3,497 participants, we

[*]Work done during a student researcher role at Google.

develop a cohort of synthetic personas to evaluate and optimize LLM-based agents. Our specific contributions include:

- Developing a Transformer-based variant of time-series Generative Adversarial Network (TimeGAN) Yoon et al. (2019) for creating synthetic time-series from the real population data that accurately reflects nuances of data missingness and population distributions while enabling privacy preservation.
- Combining synthetic time-series data with LLM-based multi-agent orchestrations to generate biometrically-aligned multimodal persona profiles/vignettes.
- Developing a framework for multi-turn conversations using the synthetic personas and evaluating persona quality for real-world plausibility using LLM-as-a-judge.

While focused on the postpartum period, our methodology provides scalable tooling for developing and evaluating personal health agents across diverse clinical and biometric applications.

2 METHODS

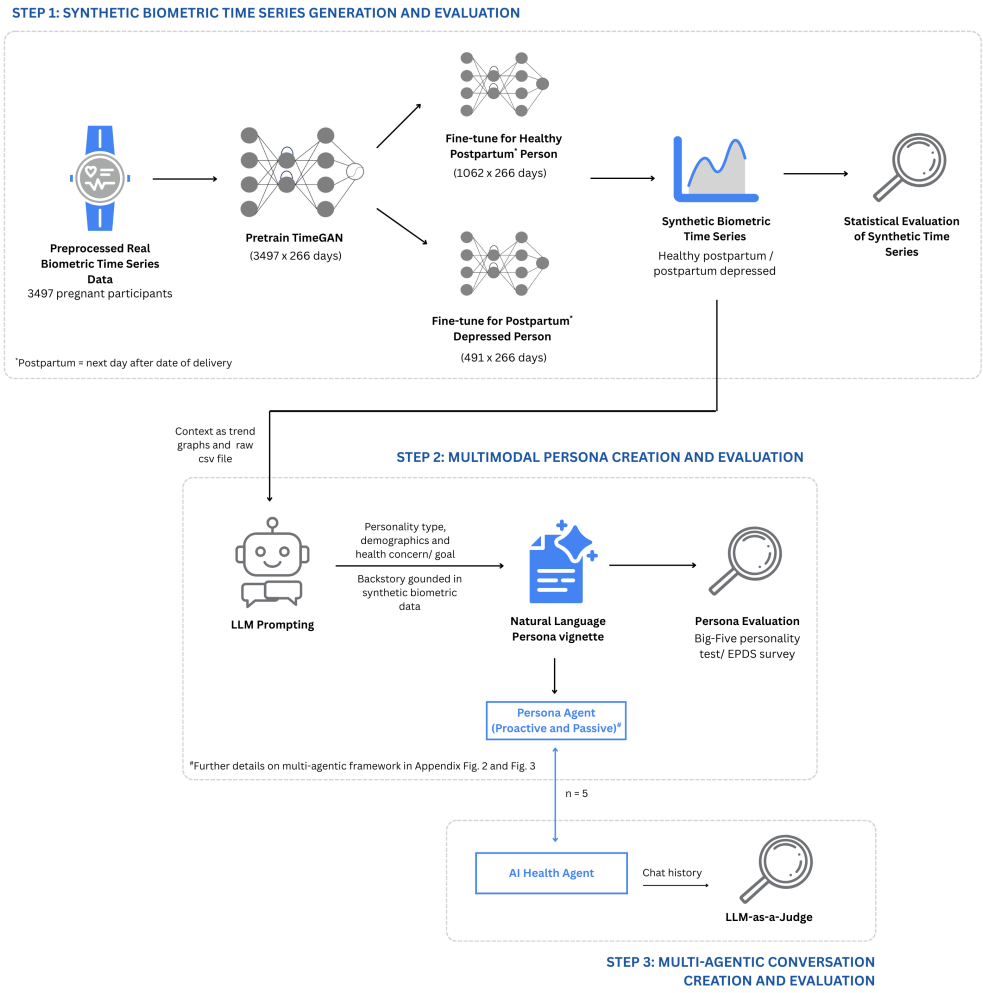

Figure 1: An overview of the multi-step methodology for generating synthetic users grounded in synthetic biometric time series data and natural language vignettes for the automated evaluation of AI health agents.

The methodology of this study is divided into three primary stages: synthetic biometric data generation using a novel Transformer-based TimeGAN architecture, the construction of multimodal synthetic personas, and the deployment of a multi-agent framework for conversational evaluation. Fig 1 describes the flow diagram for our multi step methodology.

2.1 Synthetic Time Series Data Generation and Evaluation

The generation of 29 distinct synthetic daily wearable biometrics, including Resting Heart Rate (RHR), Heart Rate Variability (HRV), sleep stages (deep, light, REM), and step count, utilizes a Transformer-based Time-series Generative Adversarial Network (TimeGAN). This approach builds on the ability of self-attention mechanisms to capture the long-range temporal dependencies inherent in the average 266-day pregnancy journey. Currently, we focus on the pregnancy biometric trends and evaluate it on the next day of delivery (postpartum day 1). To prepare the data, raw per-user daily average logs were transformed into a structured 3D tensor defined by sequences, time steps, and features. Participant sequences were filtered for a minimum 90% wear-time and truncated to the most recent 266 days to represent the average gestational period. Intra-sequence missing values were handled via a combination of forward-fill and backward-fill imputation, while sequences shorter than the target length were padded. Final features were standardized using the mean and the standard deviation of the primary training cohort to stabilize the GAN training process.

The core TimeGAN framework was implemented with Transformer blocks replacing traditional Recurrent Neural Networks (RNNs) to enhance temporal modeling. The Autoencoder, comprising an Embedder (E) and Recovery (R) network, maps high-dimensional biometric data into a latent space (Z) with a dimension of 128. This pre-training phase ensures the Generator learns to produce samples residing on a realistic manifold. The Generator (G) and Discriminator (D) similarly utilize Transformer blocks with Positional Encoding to process sequential data. To ensure the synthetic data is statistically indistinguishable from real records, we employed a Wasserstein GAN with Gradient Penalty (WGAN-GP) lossGulrajani et al. (2017). The adversarial loss is calculated as:

$$L_{adv} = -\mathbb{E}[\mathbf{D}(X)] + \mathbb{E}[\mathbf{D}(\tilde{X})] + \lambda_{GP} \cdot \text{GP} \tag{1}$$

where X represents real biometric data, $\tilde{X}$ denotes synthetic data generated by G, and $\mathbf{D}$ is the critic (discriminator). To enforce the 1-Lipschitz continuity constraint required for training stability, the Gradient Penalty (GP) is calculated as $\mathbb{E}_{\hat{X}\sim\mathbb{P}_{\hat{X}}}[(\|\nabla_{\hat{X}}\mathbf{D}(\hat{X})\|_2 - 1)^2]$, where $\hat{X}$ is sampled uniformly along straight lines between pairs of points from the real and synthetic data distributions. The weighting coefficient λ_{GP} was set to 10 for all experiments.

The model was trained in two stages: initial pre-training on a large dataset of 3,497 participants to learn fundamental biometric patterns, followed by fine-tuning on smaller target cohorts (e.g., individuals with postpartum depression and healthy postpartum) to capture specific signatures like sleep fragmentation. Hyperparameters include a latent dimension of 128 and an Adam optimizer with a learning rate of 10^{-4}.

The reliability of the synthetic output was assessed via Distributional Alignment and employed Mean Absolute Error (MAE) and Standard Deviation (SD) shift as metrics to verify that the model captured the central tendencies and inter-individual variability inherent in the real-world pregnancy and postpartum cohorts.

2.2 Persona Creation and Evaluation

The persona creation process employs a merger of two distinct data streams: Demographic Context and Time Series Context. This multimodal integration ensures the resulting vignettes are both narratively rich and biometrically grounded.

Demographic Context was generated using the Gemini 2.5 Flash Large Language Model (LLM) Comanici et al. (2025). Through an engineered prompt, the LLM acted as an expert in creating vignettes for pregnant and postpartum participants. Aligned parameters including age, ethnicity, Big-Five personality traits (Goldberg, 1992; Serapio-García et al., 2023), and health goals were generated to represent specific health barriers, such as postpartum depression, without explicitly naming the condition. These text-based vignettes were then fused with the synthetic biometric timeseries to create a complete multimodal profile. Table 1 shows one of the synthetic persona

depicting a postpartum depressed person and Table 2 shows the synthetic persona generated for healthy postpartum person.

To ensure clinical and behavioral plausibility, generated personas underwent a two-fold validation process. First, they were subjected to five rounds of condition-specific Q&A sessions (e.g., screening for depressive symptoms using (Cox & Holden, 2003)) to ensure the backstory aligned with the intended health barrier. Second, Big-Five personality tests (Goldberg, 1992) were used to verify adherence to persona's defined psychological profile.

2.3 Multi-Agent Framework for Conversation and Evaluation

Interactions between the synthetic personas and an AI Health Bot were facilitated using the `Autogen` library Wu et al. (2024), with all agents powered by the `gemini-2.5-flash` model.

2.3.1 Multi-Agent Conversation Architectures

The framework utilizes two distinct architectures. The *passive framework* mirrors user-initiated interactions and the *proactive framework* assesses the AI Health Agent's ability to initiate dialog based on biometric trends.

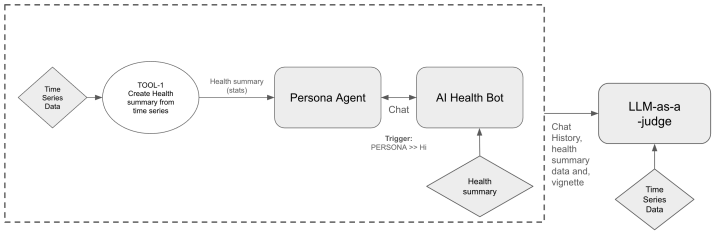

Figure 2: Multi-agentic Passive Persona Framework

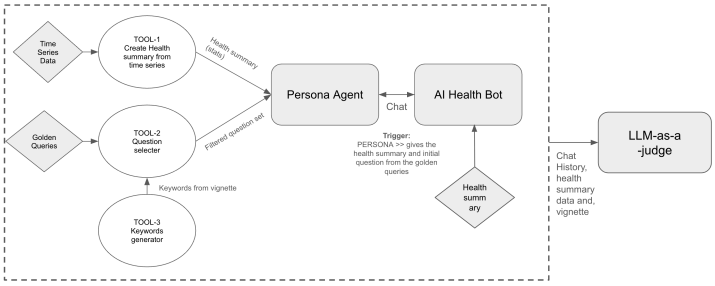

Figure 3: Multi-agentic Proactive Persona Framework

The proactive framework (Fig. 2) assesses the AI Health Agent's ability to initiate dialog based on biometric trends. The AI agent was instructed to formulate a compassionate opening line acknowledging specific data trends. Conversely, the Persona Agent was initialized with an "apathetic" system message, providing minimal input (starting with "Hi") to force the AI agent to drive the conversation and identify health barriers over 5 turns.

The passive framework (Fig. 3) mirrors user-initiated interactions where the user provides specific queries. An automated tool parsed the persona vignette to extract about 10 unique keywords, which were then used to filter a "golden query set" of postpartum questions. This produced a relevant question list string injected into the Persona Agent's system message. The Persona Agent initiated the dialog by combining a biometric health summary with the first relevant question, engaging the AI Health Bot for a maximum of 5 turns.

2.3.2 LLM-as-a-Judge Evaluation

Following the simulations, an LLM-as-a-Judge agent performed a granular assessment of the dialog. The judge applied seven metrics: Factual Correctness & Data Usage, Context Relevance,

Plausibility, Realism, Persona Adherence, Coherence, and Goal Progression. The judge provided independent scores (1–5) and qualitative reasoning for both the Persona Agent and the AI Health Bot to identify specific performance gaps in health-informed dialog.

3 RESULTS

The evaluation of our framework focused on two primary dimensions: the statistical analysis of the synthetic biometric data, and the effectiveness of the multi-agent conversational framework in simulating realistic health interactions.

3.1 EVALUATION OF SYNTHETIC TIME SERIES GENERATION

To test how well our TimeGAN model works, we compared the average values of the generated data against the real data. For the healthy postpartum cohort, the model accurately captured the averages for biometric measurements like Total Step Count, Resting Heart Rate, HRV, etc. In the postpartum depressed cohort, the generated data showed a very close match in sleep patterns, specifically, the average for Deep Sleep (66.16 min) was nearly identical to the real average (65.97 min). Across both groups, even though the generated data had slightly less variety than the real data, the consistent Relative Error levels show that the TimeGAN data could be a reliable substitute for real sensor data. Figures 4 and 5 present the mean and standard deviation for the healthy postpartum and postpartum depressed cohorts respectively, comparing real-data distributions against synthetically generated data.

Furthermore, to visualize the high-dimensional cross-cohort data, we utilized the t-Distributed Stochastic Neighbor Embedding (t-SNE) Cieslak et al. (2020) implementation from the `scikit-learn` library. As shown in Figure 6, the clear separation between synthetic postpartum healthy and depressed clusters demonstrates that the framework successfully captured distinct longitudinal signatures.

3.2 MULTI-AGENT CONVERSATION EVALUATION

The generated personas were integrated into a multi-agent framework to evaluate interaction quality between synthetic users and the AI Health Bot. An LLM-as-a-Judge agent provided granular scores (1–5) across key metrics, as detailed in Table 3 and Table 4. These results indicate high performance in Persona Adherence and Plausibility in successful trials, while also identifying critical technical failures. The AI Health Bot demonstrated a high capacity for data-driven opening statements in "Proactive" mode, identifying health trends without user prompting. In "Passive" mode, the bot successfully addressed 10 unique keywords related to the persona's specific health concerns, demonstrating robust contextual awareness.

4 DISCUSSION AND FUTURE WORK

This work introduces a multi-agent framework for creating multimodal synthetic personas designed to serve as a standardized, privacy-preserving environment for benchmarking future conversational health agents. By integrating narrative vignettes with longitudinal synthetic biometric time-series data, our approach offers a robust testing ground for assessing how agents handle the complex, physiologically dynamic shifts of the pregnancy period. While our current Transformer-based TimeGAN successfully models central population tendencies, the high statistical similarity between generated samples highlights a need for future work to increase generative diversity, ensuring agents are tested against biological outliers and rare clinical scenarios. Additionally, while the LLM-as-a-Judge serves as a scalable initial evaluation metric, human validation from OB/GYNs and psychologists remains essential to fully verify medical safety and therapeutic empathy in agent interactions. We intend to expand this persona library to include diverse socioeconomic backgrounds and comorbidities, further strengthening the generalizability of our framework as a stress-testing tool. Ultimately, this methodology addresses the monomodal constraints of existing persona research and provides a scalable way to identify performance gaps, offering a framework for the development of safer, data-informed health agents in the postpartum period and beyond.

REFERENCES

Matthew C Cieslak, Ann M Castelfranco, Vittoria Roncalli, Petra H Lenz, and Daniel K Hartline. t-distributed stochastic neighbor embedding (t-sne): A tool for eco-physiological transcriptomic analysis. *Marine genomics*, 51:100723, 2020.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

John Cox and Jeni Holden. *Perinatal mental health: A guide to the edinburgh postnatal depression scale (EPDS).* Royal College of Psychiatrists, 2003.

Poorvesh Dongre. Physiology-driven empathic large language models (emllms) for mental health support. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–5, 2024.

Lewis R Goldberg. The development of markers for the big-five factor structure. *Psychological assessment*, 4(1):26, 1992.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J Taylor, and Dan Roth. Know me, respond to me: Benchmarking llms for dynamic user profiling and personalized responses at scale. *arXiv preprint arXiv:2504.14225*, 2025.

Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. Health-llm: Large language models for health prediction via wearable sensor data. *arXiv preprint arXiv:2401.06866*, 2024.

Vahid Rahimzadeh, Erfan Moosavi Monazzah, Mohammad Taher Pilehvar, and Yadollah Yaghoobzadeh. Synthia: Synthetic yet naturally tailored human-inspired personas. *arXiv preprint arXiv:2507.14922*, 2025.

Zhiwei Ren, Junbo Li, Minjia Zhang, Di Wang, Xiaoran Fan, and Longfei Shangguan. Toward sensor-in-the-loop llm agent: Benchmarks and implications. In *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems*, pp. 254–267, 2025.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. Personagym: Evaluating persona agents and llms. *arXiv preprint arXiv:2407.18416*, 2024.

Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. Personality traits in large language models. 2023.

Xiaoyang Wang, Hongming Zhang, Tao Ge, Wenhao Yu, Dian Yu, and Dong Yu. Opencharacter: Training customizable role-playing llms with large-scale synthetic personas. *arXiv preprint arXiv:2501.15427*, 2025.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First conference on language modeling*, 2024.

Jinsung Yoon, Daniel Jarrett, and Mihaela Van der Schaar. Time-series generative adversarial networks. *Advances in neural information processing systems*, 32, 2019.

Taedong Yun, Eric Yang, Mustafa Safdari, Jong Ha Lee, Vaishnavi Vinod Kumar, S Sara Mahdavi, Jonathan Amar, Derek Peyton, Reut Aharony, Logan Douglas Schneider, et al. Sleepless nights, sugary days: Creating synthetic users with health conditions for realistic coaching agent interactions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14159–14181, 2025.

A APPENDIX

Table 1: Comprehensive Clinical Persona: Elena Rossi

Category	Details
Persona Type	postpartum Depressed (next day of delivery)
Name	Elena Rossi
Demographics	34 yrs, 70 kg, 150 cm, Female, Latina
Clinical Context	Term Delivery, Gestational Anxiety
Occupation	Marketing Executive (on Parental Leave)
Personality Type	Neuroticism (Big-Five)
General Demeanor	Tearful, high-anxiety, expressing deep guilt and inadequacy.
Digital Literacy	Proficient with phone interfaces; Not new to Fitbit.
Health Scenario	*Activity:* Physiological Stress and Poor Sleep. *Sleep:* Persistent increase in RHR and decrease in total sleep duration throughout pregnancy.
Concern & Goal	*Concern:* Extreme fatigue and non-stop feeling of dread/overwhelm. *Goal:* Decrease average daily RHR by 5 BPM within one week using 10-minute relaxation techniques.
Backstory	Elena, a former Marketing Executive, gave birth to her second child after a term delivery. As a long-term Fitbit user, she tracked a continuous increase in her Resting Heart Rate and a drop in sleep throughout her pregnancy, signaling chronic stress. Now one day postpartum, her history of Neuroticism and anxiety has made this experience much harder than her first. She feels overwhelmed and uses her high knowledge level to self-criticize, feeling like a failure because she cannot "snap out of it." She is currently using wearable data to confirm her body is in a state of high alert and aims to use tiny habits, like deep breathing, to calm the internal "wiredness" preventing her from resting and bonding with her baby.

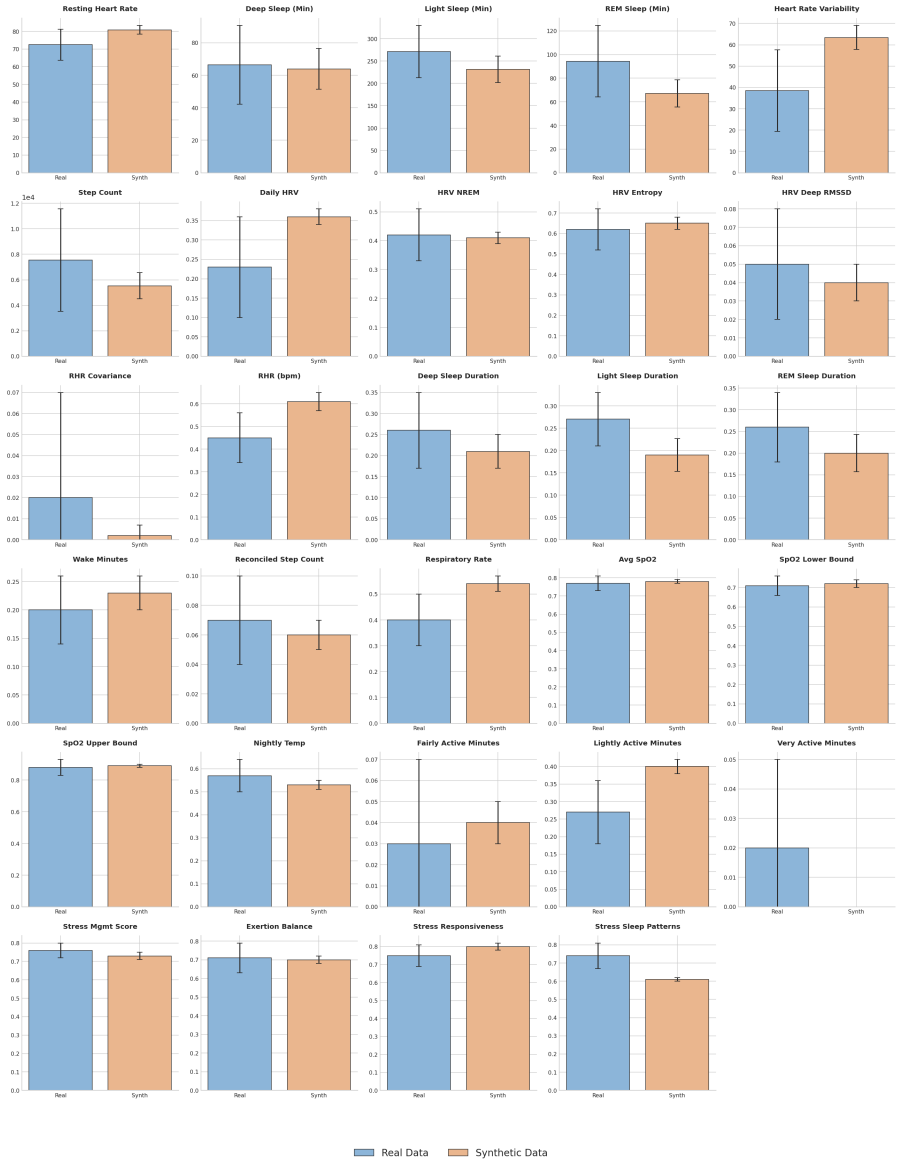

Figure 4: Comparison of real vs. synthetic wearable biometrics for the healthy postpartum cohort, showing mean with standard deviation error bars

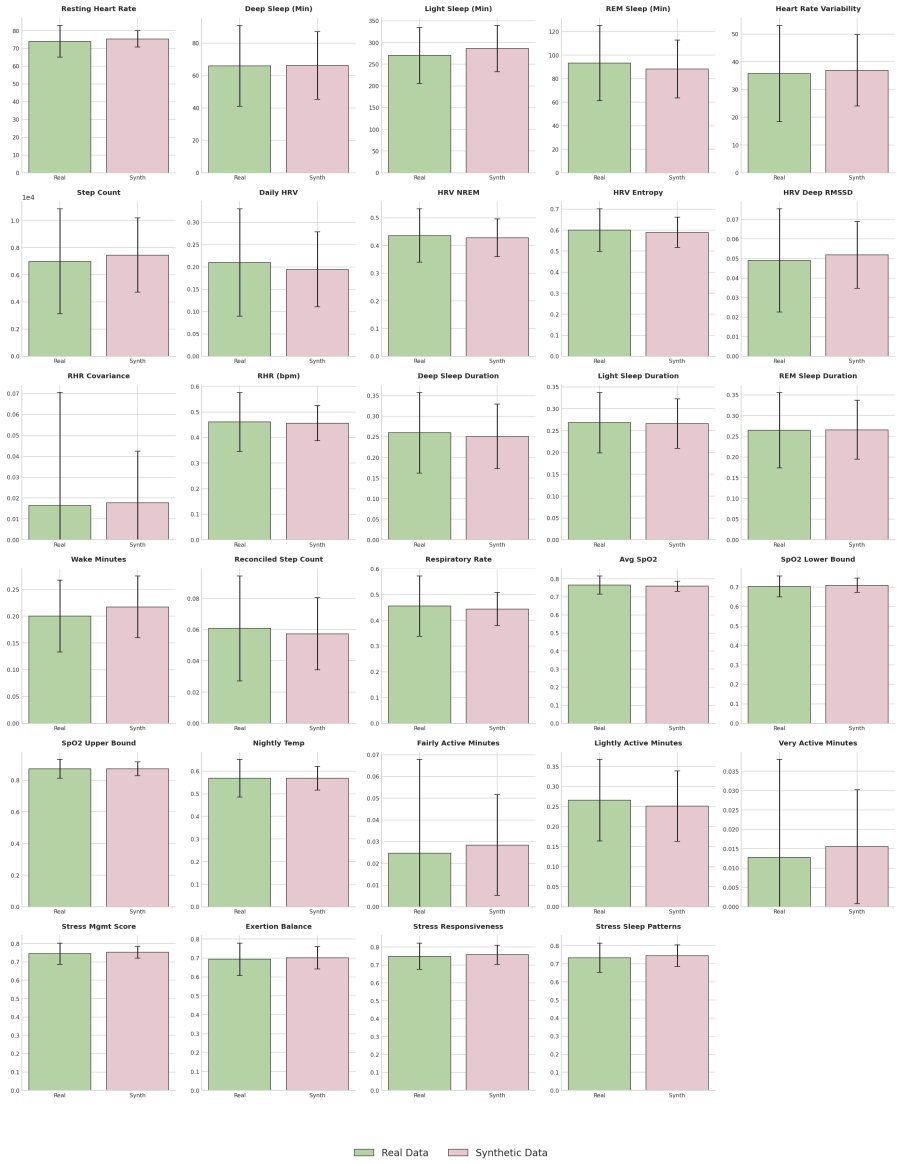

Figure 5: Comparison of real vs. synthetic wearable biometricsfor the postpartum depressed cohort, showing mean with standard deviation error bars

Table 2: Clinical Persona Profile: Healthy Postpartum Case Study

Category	Details
Persona Type	Healthy postpartum (next day of delivery)
Name	Anya Sharma
Demographics	32 yrs, 80 kg, 155 cm, Female, South Asian
Clinical Context	Term Delivery, No underlying conditions
Occupation	UX Designer (on Parental Leave)
Personality Type	Conscientiousness (Big 5)
General Demeanor	Optimistic but exhausted; highly motivated for wellness.
Digital Literacy	Very proficient with phone interfaces; Not new to Fitbit.
Health Scenario	*Sleep & Recovery:* Sudden, dramatic fragmentation of rest dictated by newborn feeding schedule. *Metrics:* Heart rate metrics successfully accommodated pregnancy; robust health history.
Concern & Goal	*Concern:* Sleep fragmentation inhibiting critical RHR drop and recovery. *Goal:* Maximize 20–30 minute restorative rest bouts and achieve 10 minutes of slow walking per day.
Backstory	Anya, a first-time mother one day post-delivery, is experiencing profound joy mixed with the physical toll of labor. Her pregnancy was clinically uneventful, and she successfully monitored her metrics throughout. However, her predictable schedule has dissolved into an unpredictable sleep cycle, managing only short bursts of rest interrupted by the baby's needs. Her current focus has shifted from high activity to healing and recovery. She aims to use the AI coach to help her embrace "micro-breaks" for rest and ensure she completes gentle, doctor-prescribed movements, such as walking to the nursery, to help her body transition out of the high cardiovascular demand of pregnancy.

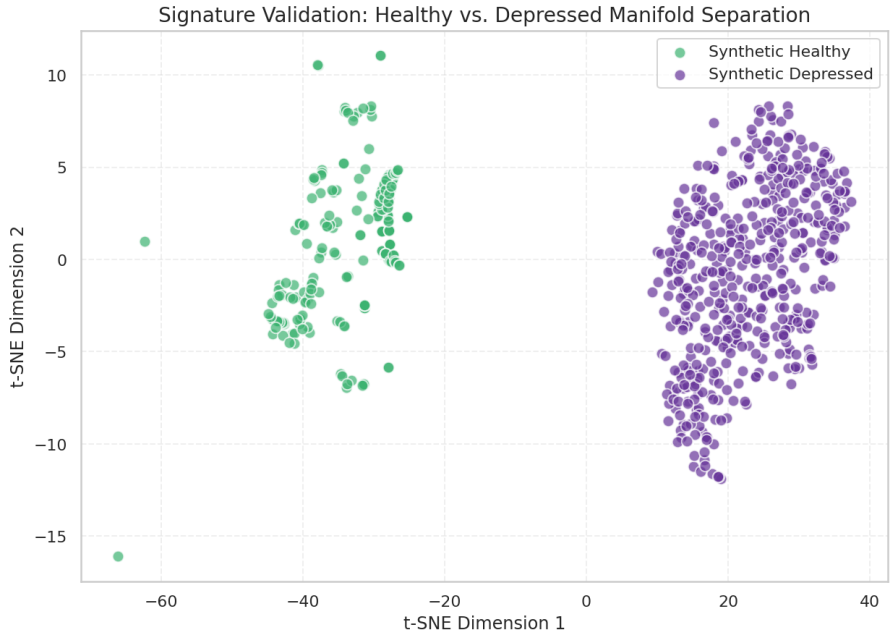

Figure 6: Signature Validation via t-SNE Manifold.

Table 3: Evaluation of the conversation between proactive AI agent and the persona by LLM-as-a-judge

Category	Score	Detailed Reasoning & Evidence
Factual Correctness	5/5	Accurately interpreted RHR (84-86 bpm) and Total Sleep (6.76h). Correctly linked elevated RHR to fatigue and emotional strain.
Context Relevance	5/5	Revolved around core themes of fatigue and apathy. AI suggested "passive, environmental shifts" matching the user's goal for non-demanding rest.
Plausibility	5/5	Persona's responses ("Energy feels... flat.") were a perfect representation of apathetic PPD. AI's gentle approach felt believable for health tech.
Realism	5/5	Interactions were detailed without being robotic. Brief emotional tone was consistent with the persona's state.
Persona Adherence	5/5	**Persona:** Portrayed extreme flatness/anxiety. **AI Health Bot:** Maintained compassionate, data-informed role.
Coherence	Yes	Logical progression from initial empathy to a clear next-step proposal and termination.
Final Summary	This conversation was exceptional, with the AI skillfully integrating health data to provide highly relevant and compassionate support that perfectly matched the client's apathetic persona.	
OVERALL RATING	**5 / 5**	

Table 4: Evaluation of the conversation between passive AI agent and the persona by LLM-as-a-judge

Criterion	Score	Key Reasoning
Factual Correctness	4/5	Accurately used RHR (84.45) and Sleep (6.76h). Factually correct on PPD/PPA, though "HR Cov" was not utilized.
Context & Relevance	5/5	Perfectly aligned with persona themes of fatigue and apathy; advice linked directly to symptoms and data.
Plausibility	2/5	Undermined by a technical glitch where the AI repeated a lengthy response verbatim in Turn 4.
Realism	1/5	AI responses were overly verbose; verbatim repetition across two turns made the conversation feel dysfunctional.
Persona Adherence	2/5	**Persona:** Persona became too verbose in Turn 5. **AI Health Bot:** AI failed to maintain a coherent role due to the verbatim repetition error;
Coherence	No	Duplication of AI response in Turn 4 created a severe break in logic and halted goal progression.
Summary	Effectiveness was critically undermined by a severe technical error where the AI repeated a response verbatim. Despite factually sound content, the flaw made the interaction unrealistic.	
Overall Rating	**2 / 5**	