# Personalized Remote Health Monitoring and Clinical Report Generation Using Multi-Agent LLM Framework on Wearable Data

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Remote health monitoring with consumer wearables can continuously capture sleep, activity, and cardiorespiratory signals, but raw metrics are difficult to translate into clinically meaningful guidance. We present a modular multi-agent large language model (LLM) framework that converts longitudinal wearable data plus personal metadata (age, BMI, smoking status, and comorbidities) into clinician-style daily and weekly narratives with targeted recommendations. Our pipeline comprises specialized agents for (i) anomaly detection, (ii) trend summarization, (iii) lifestyle suggestions, (iv) final daily synthesis, and (v) weekly aggregation. To enable reproducible evaluation without PHI, we generate a medically grounded synthetic dataset of *50* users over *7* days, with anomalies injected in *35%* of users. Each user contributes 2,016 five-minute physiological samples, 1,008 ten-minute step intervals, and 224 sleep stage entries per week, yielding more than 160,000 structured records across the cohort. On injected anomalous events, the Anomaly Agent achieves precision 0.89 and recall 0.91. In a blinded expert review (two clinicians, one public-health researcher) of system outputs, average ratings (1–3 scale) were: clinical relevance 2.7, personalization 2.5, and clarity 2.8. An ablation shows that removing personalization reduces individualization by $\sim$24% ($\approx$25%), and replacing the multi-agent pipeline with a single-agent baseline lowers average scores across relevance, personalization, and clarity. **Code & data:** `https://github.com/AliArshadswl/Agents4Science`.

## 1 Introduction

The rising prevalence of chronic diseases, driven by aging populations and escalating healthcare needs, underscores the need for scalable, cost-effective health monitoring. Non-communicable diseases (NCDs) account for about 38 million deaths annually, or 63% of global mortality [1, 2]. Multimorbidity, defined as having two or more chronic illnesses, affects over half of adults aged 65+ [2], creating major strain on healthcare systems, particularly in underserved regions.

Wearable devices such as smartwatches, fitness trackers, and biosensor patches enable continuous, passive collection of heart rate, oxygen saturation ($SpO_2$), heart rate variability (HRV), activity, and sleep. Prior studies highlight their potential for chronic disease management and patient engagement [3, 4]. For older adults, large-scale surveys show that wearable use can improve physical health outcomes, especially among those aged 70+ or living with children [5].

However, the clinical utility of wearables is limited by raw data streams that are difficult for patients and caregivers to interpret. Numeric dashboards often fail to provide personalized, actionable insights, leading to confusion or inaction.

Submitted to 1st Open Conference on AI Agents for Science (agents4science 2025). Do not distribute.

Recent work shows that large language models (LLMs) can help bridge this gap. The Personal Health Insights Agent (PHIA) demonstrated accurate interpretation of wearable behavioral data [6], while LLM-based pipelines have improved tasks such as heart rate estimation from photoplethysmography (PPG) [7]. Multimodal systems like LLaSA [8] and SensorLM [9] extend LLM reasoning to diverse sensor streams. In parallel, multi-agent AI frameworks—where specialized agents collaborate with reasoning and memory—are gaining traction for healthcare tasks including chronic disease management and elder care [10, 11].

Together, these findings support three central claims: (i) wearables generate valuable continuous physiological data; (ii) without contextual interpretation, this data is underutilized; and (iii) LLM-based multi-agent systems offer a promising pathway to generate clinically meaningful, user-friendly health reports.

# 2 Related Work

## 2.1 Wearable Health Monitoring: From Data Collection to Clinical Insights

Consumer wearables have transformed personal health monitoring by continuously capturing signals such as heart rate variability, skin temperature, blood oxygen saturation ($SpO_2$), electrodermal activity, and movement patterns [12, 13]. Despite this richness, commercial platforms still rely on simple analytics such as static thresholds, basic trends, and population norms, which limit personalization and clinical utility [14, 15].

Machine learning methods have advanced physiological time-series analysis. Deep learning models, particularly CNNs and RNNs, have achieved strong results in stress detection [16, 17], arrhythmia classification [18], and sleep stage prediction [19, 20]. Multimodal fusion combining PPG, accelerometry, and thermography has further improved robustness [21–23]. Yet these models are often black boxes, with limited interpretability and little contextual reasoning, hindering clinical adoption [24, 25].

## 2.2 Large Language Models in Healthcare: Bridging Text and Physiological Data

Transformer-based LLMs have redefined natural language understanding [26]. Foundation models such as BERT [27] and GPT [28, 29] inspired biomedical variants like BioBERT [30], ClinicalBERT [31], and BioGPT [32]. These models, pre-trained on biomedical corpora, support tasks such as medical question answering and record summarization [33, 34]. More recent instruction-tuned systems extend their reach: Med-PaLM performs at near-expert level on licensing exams [35], and ClinicalGPT and ChatDoctor show diagnostic reasoning capabilities [36, 37].

Despite these advances, biomedical LLMs mainly process textual or structured inputs and cannot directly interpret high-frequency physiological data from wearables [39, 40]. Closing this gap requires representational methods that let LLMs reason contextually and interpret biosignals in clinically relevant ways [41, 42].

## 2.3 Multi-Agent Systems: Distributed Intelligence for Complex Health Tasks

Comprehensive health monitoring benefits from decomposing tasks into specialized components. Multi-agent systems distribute responsibilities across agents for anomaly detection, trend summarization, or recommendations [43, 44]. Recent frameworks such as AutoGen [45], AgentVerse [46], and CrewAI [47] show that role-specific collaboration can improve coherence and robustness [48, 49].

For wearable health monitoring, multi-agent setups mirror healthcare teams—triage nurses assessing risks, physicians reasoning diagnostically, wellness coaches suggesting lifestyle changes. This approach produces interpretable, clinically aligned narratives and allows modular improvement and integration of domain expertise [52–54].

## 2.4 Research Gap and Contribution Positioning

Progress in wearable sensing, biomedical AI, and language modeling has been rapid, yet these fields remain disconnected. Wearable platforms still lack contextual intelligence for clinical decision

support, while biomedical LLMs are limited to textual data and cannot directly interpret continuous physiological signals [65, 66]. Multi-agent LLM systems provide a natural convergence point, offering modular, interpretable, and clinically aligned reasoning over biosignal data [52].

We address this gap with a multi-agent LLM framework that links continuous physiological monitoring to contextual language understanding. By integrating personal metadata (age, BMI, comorbidities) and using a synthetic dataset for evaluation, our approach improves personalization, interpretability, and reproducibility [67, 68]. This work represents a step toward intelligent, automated health monitoring systems that deliver clinically relevant insights from wearable data [69, 70].

# 3 Dataset Creation Methodology

We generated a synthetic wearable health dataset using evidence-based medical modeling to approximate realistic physiological patterns. The framework integrates epidemiological statistics, clinical guidelines, and literature-based multipliers to ensure both medical plausibility and population representativeness.

## 3.1 Population and Risk Factors

We simulated 50 adults (ages 25–80) with gender-specific anthropometrics derived from NHANES data [71]. BMI distributions matched U.S. population norms: 32% normal, 36% overweight, and 32% obese [72]. Disease prevalence was age- and BMI-stratified, with risks for diabetes, hypertension, and cardiovascular disease modeled from established sources [73, 75, 76]. Individual health profiles combined these factors with smoking status and comorbidities.

## 3.2 Fitness and Activity Modeling

A composite fitness score incorporated age, BMI, and comorbidities to classify users into high, moderate, or poor fitness. Step counts followed age-based pedometer baselines [78], adjusted by fitness level and disease burden (e.g., heart disease reducing activity by 30–60%, COPD by 20–50%). Circadian patterns reflected structured activity in high-fitness users and irregular rhythms in those with poor health [84].

## 3.3 Sleep and Physiological Signals

Sleep architecture was modeled with age-related declines in deep sleep and wake after sleep onset [85], plus disease-specific disruptions (e.g., sleep apnea, COPD) [86]. Cardiovascular parameters included resting heart rate, blood pressure, and HRV adjusted for age, fitness, and disease status [88, 89]. Respiratory signals ($SpO_2$, rate) and glucose dynamics were simulated with normal and pathological patterns, including dawn phenomenon in diabetes [73].

## 3.4 Anomalies and Events

To emulate clinical variability, anomalies were injected into 35% of users. Modeled events included arrhythmias, hypertensive crises, hypoxemia, and hyperglycemia, with frequencies informed by hospitalization data. Each "bad day" reflected elevated risks based on comorbidities (e.g., +18% in COPD, +12% in heart disease).

## 3.5 Data Quality and Constraints

Realistic sensor noise and missingness were added (e.g., 91% $SpO_2$ availability, 95% step counter reliability). Physiological bounds (heart rate 30–220 bpm, glucose 40–600 mg/dL, $SpO_2$ 80–100%) and inter-parameter correlations were enforced to maintain coherence [**?** ].

## 3.6 Dataset Scale

Each user contributed 2,016 five-minute physiological samples, 1,008 ten-minute activity intervals, and 224 sleep stage entries per week, yielding over 160,000 structured records across the cohort.

Aggregate metrics matched population surveys: mean daily steps 6,847 (SD 3,421), with distributions of 30% high, 40% moderate, and 30% poor fitness [92].

# 4 Methodology

We developed a communicative multi-agent LLM framework that interprets wearable data streams and personal metadata to generate structured daily and weekly health reports. The system is designed to support both individual self-monitoring and remote clinical assessment, and emphasizes collaboration among agents to improve coherence and reliability.

## 4.1 Motivation and Design Principles

Wearable devices provide continuous physiological measurements such as heart rate variability (HRV), oxygen saturation ($SpO_2$), sleep, and physical activity. However, these data streams are difficult to interpret directly, particularly for individuals with chronic conditions or their caregivers. Our system is guided by three principles: clinical grounding through alignment with medical literature and guidelines; personalization by incorporating demographic and medical metadata; and modular scalability through a distributed, stateless architecture suitable for cloud or edge deployment.

## 4.2 System Architecture

The architecture comprises three stages: data ingestion, collaborative agent analysis, and report generation. Inputs include wearable sensor data and personal metadata (age, sex, BMI, smoking status, and comorbidities). Five specialized agents process these inputs:

- Summary Agent: produces a daily wellness overview.
- Anomaly Agent: identifies deviations from medically informed thresholds, adapted to user-specific risk profiles.
- Suggestion Agent: generates personalized lifestyle and preventive recommendations.
- Final Summary Agent: integrates the outputs of other agents into a coherent daily narrative.
- Weekly Agent: aggregates reports across seven days, identifies longitudinal trends, and highlights recurring anomalies.

Agents do not operate in isolation but communicate iteratively. For instance, the Anomaly Agent may share alerts with the Summary Agent, prompting revisions, while the Suggestion Agent adapts recommendations if anomalies recur. The Final Summary Agent performs consensus-building, ensuring consistency across agent outputs. This communication is facilitated by a shared backend using the DeepSeek API, which supports structured message exchange among agents.

## 4.3 Personalized Prompting and Communication

Prompts are dynamically constructed to encode both structured physiological data (e.g., average HRV, minimum and maximum $SpO_2$, total sleep time) and descriptive metadata (e.g., age, BMI, comorbidities). Intermediate agent outputs are also exchanged as part of the prompt context, allowing agents to refine their reasoning collaboratively. This inter-agent communication improves contextual interpretation and reduces inconsistencies across anomaly detection, recommendations, and final summaries.

## 4.4 Pipeline Execution and Output

Each day begins with preprocessing of wearable data into a standardized prompt format. The Summary, Anomaly, and Suggestion agents operate in parallel but exchange intermediate results. Their outputs are synthesized by the Final Summary Agent into a narrative accessible to patients and clinicians. After seven days, the Weekly Agent reviews temporal patterns, incorporates recurring anomalies, and produces a consolidated report, which is rendered as a LaTeX-based PDF.

### 4.5 Deployment and Integration

The system is containerized, with each agent implemented as a stateless service to support scalability and distributed execution. Inter-agent communication is lightweight, enabling deployment both in cloud-based platforms and at the network edge. While the DeepSeek API served as the language model backend in this study, the architecture is model-agnostic and can incorporate alternative or fine-tuned LLMs. Future extensions include direct integration with Apple HealthKit, Fitbit APIs, and electronic medical record (EMR) systems, enabling secure ingestion of real-world data and sharing of reports with healthcare providers.

In summary, the methodology demonstrates how communicative multi-agent reasoning, combined with personalized prompting and DeepSeek-based report generation, can transform raw wearable data into clinically relevant and interpretable health summaries.

## 5 Evaluation and Usefulness

To assess the effectiveness, personalization quality, and clinical relevance of our multi-agent LLM-based health monitoring system, we employed a four-pronged evaluation strategy: (1) anomaly detection validation on synthetic datasets, (2) structured clinical expert review, (3) LLM-as-a-judge scoring using the ChatGPT-5 API, and (4) ablation-based robustness and baseline comparison. This combined approach compensates for the absence of gold-standard clinical labels by triangulating between controlled synthetic data, domain expertise, and standardized LLM-based evaluations.

### 5.1 Design of Evaluation Metrics

Evaluating multi-agent clinical summarization systems is challenging because no absolute "ground-truth" reports exist. To address this, we developed a structured rubric emphasizing both clinical safety and patient-centered communication. The rubric was co-designed with two physicians and a public health researcher, then refined in a small pilot review.

The final evaluation framework consisted of five dimensions:

1. **Report Appropriateness** — whether summaries focus on clinically meaningful findings without trivializing normal variation or exaggerating minor fluctuations.
2. **Individualization to Case** — whether outputs adapt to user-specific demographics and risk factors (e.g., age, BMI, smoking, comorbidities).
3. **Safety and Risk Feasibility** — whether acute risks are flagged in a responsible and feasible manner (e.g., not over-escalating benign cases).
4. **Quality of Justification** — whether flagged anomalies are explained with reasoning or supporting evidence, rather than simply listed.
5. **Clarity and Accuracy** — whether reports use medically correct definitions while remaining understandable to laypersons and caregivers.

Each dimension was scored on a 3-point Likert scale (1 = Poor, 2 = Adequate, 3 = Excellent). Human raters were provided with written guidelines and illustrative examples (e.g., "adequate justification" = reporting desaturation without causal reasoning; "excellent justification" = linking desaturation to concurrent smoking status and BMI).

For reliability, inter-rater agreement among experts was calculated, yielding Cohen's $\kappa = 0.71$ (substantial agreement). The LLM-as-a-judge (ChatGPT-5) was prompted with the same rubric and rating guidelines to ensure comparability.

### 5.2 Anomaly Detection Validation

In the synthetic dataset, 35% of users were assigned injected anomalies such as reduced HRV (<40ms), nocturnal $SpO_2$ desaturations (<90%), or elevated resting heart rate. These anomalies served as pseudo ground-truth.

The Anomaly Detection Agent correctly identified 41 of 45 injected anomalies, with precision $= 0.89$ and recall $= 0.91$. Most errors occurred in borderline $SpO_2$ events near the 90% cutoff. This demonstrates sensitivity to meaningful physiological deviations while maintaining relatively low false-positive rates.

### 5.3 Clinical and LLM-as-a-Judge Review

We then evaluated the interpretability and usefulness of generated summaries using both clinical experts and LLM-as-a-judge scoring.

- **Human Reviewers:** Two physicians and one public health researcher reviewed 15 daily summaries (10 anomalous, 5 non-anomalous).
- **LLM-as-a-Judge:** ChatGPT-5 was independently prompted with the same rubric and guidelines to generate standardized ratings.

**Results.** Table 1 reports the average scores (scale 1–3).

Table 1: Comparison of human expert vs. LLM-as-a-judge ratings (average / 3)

| **Criterion** | **Human Experts** | **LLM-as-a-Judge** |
|---|---|---|
| Report Appropriateness | 2.7 | 2.8 |
| Individualization | 2.5 | 2.6 |
| Safety and Risk Feasibility | 2.6 | 2.5 |
| Quality of Justification | 2.4 | 2.3 |
| Clarity and Accuracy | 2.8 | 2.9 |

Both human and LLM reviewers found strengths in clarity and appropriate risk highlighting, particularly for elderly users with multi-factor risks (e.g., low $SpO_2$ combined with smoking). Human reviewers tended to be more critical of weak justifications, while the LLM was slightly more lenient but highly consistent. The alignment across methods supports the feasibility of using LLM-as-a-judge to complement limited expert availability.

### 5.4 Ablation Study: Modularity and Personalization

To quantify the contributions of different system modules, we conducted ablations on 10 users:

1. **No Personalization Inputs** (age, BMI, lifestyle),
2. **No Anomaly Detection Agent**,
3. **No Final Summary Agent (fusion)**.

Ratings (averaged across human and LLM judges) are shown in Table 2.

Table 2: Impact of ablation on summary quality (average rating / 3)

| **Model Variant** | Appropriateness | Individualization | Safety | Justification | Clarity |
|---|---|---|---|---|---|
| Full Multi-Agent System | 2.7 | 2.5 | 2.6 | 2.4 | 2.8 |
| No Personalization | 2.3 | 1.9 | 2.3 | 2.2 | 2.7 |
| No Anomaly Agent | 2.2 | 2.4 | 2.0 | 2.1 | 2.6 |
| No Final Summary Agent | 2.1 | 2.3 | 2.2 | 2.0 | 2.1 |

Disabling personalization sharply reduced case-specific tailoring, especially in chronic disease cases. Without the anomaly agent, subtle risks were underreported. Removing the final fusion agent resulted in fragmented, sometimes contradictory outputs, highlighting the necessity of multi-agent synthesis.

### 5.5 Prompt Robustness and Generalization

To test robustness, we introduced noise into physiological signals (e.g., HRV spikes, missing $SpO_2$ samples). Both human experts and the LLM judged summaries as stable, with less than 12% deviation in overall scores. This demonstrates resilience to realistic wearable data artifacts.

### 5.6 Comparison with Single-Agent Baseline

Finally, we compared our multi-agent framework against a single-agent baseline (Mistral via Ollama) that directly summarized all inputs in one prompt. On 10 users (5 anomalous, 5 non-anomalous), the multi-agent system consistently outperformed the baseline (Table 3).

Table 3: Multi-agent vs. single-agent framework (average rating / 3)

| **System** | Appropriateness | Individualization | Safety | Justification | Clarity |
|---|---|---|---|---|---|
| Multi-Agent LLM (ours) | 2.7 | 2.5 | 2.6 | 2.4 | 2.8 |
| Single-Agent Baseline | 2.2 | 1.8 | 2.1 | 1.9 | 2.3 |

Experts noted that the baseline often produced generic summaries, overlooked anomalies, and failed to adjust for user context. The modular multi-agent approach produced more focused, clinically aligned, and individualized reports—mirroring interdisciplinary medical workflows.

### 5.7 Implementation and Reproducibility

All experiments were run on Ubuntu 22.04 with an AMD Ryzen 9 7950X CPU (16 cores), 64 GB RAM, and an NVIDIA RTX 4090 GPU (24 GB VRAM). The framework was implemented in Python 3.11 with LangChain 0.2.7, Docker containerization, and the DeepSeek API as the LLM backend. Average wall-clock time was 35–40 seconds per daily report, 5 minutes per 7-day report, and under 3 hours for the full dataset of 50 users (parallelized). Synthetic datasets were generated with NumPy and Pandas; random seeds were fixed for reproducibility. Hyperparameters were held constant: temperature 0.2, max tokens 1,024, top_p 0.9. All evaluation prompts are available in the project repository for full reproducibility.

### 5.8 Code and Data Availability

To ensure transparency and reproducibility, we provide open access to all code, configuration files, and the full synthetic dataset at: `https://github.com/AliArshadswl/Agents4Science`. The repository includes instructions for dataset generation, environment setup, and report reproduction, along with example outputs for selected users. Docker images with pre-configured dependencies are also provided to facilitate straightforward replication.

### 5.9 Summary of Findings

The evaluation results demonstrate the feasibility of multi-agent LLMs for health monitoring and personalized reporting. Our system achieved high sensitivity in detecting synthetic anomalies, produced clinically relevant and individualized summaries, and showed robustness to noise in wearable-like data. Importantly, the integration of both human experts and LLM-as-a-judge provided converging evidence that reports were generally appropriate, safe, and understandable, with modest but correctable gaps in justification detail.

By combining expert-informed anomaly injection, structured rubric-based scoring, and modular ablation analysis, we established a transparent and reproducible evaluation framework that preserves user privacy while capturing clinical utility. These findings support the viability of LLM-driven summarization in digital health contexts where ground-truth annotations are scarce.

## 6 Discussion and Future Work

Our findings suggest that multi-agent LLM frameworks can reliably generate personalized health summaries from wearable-derived signals. Incorporating demographic and comorbidity metadata improved contextualization, while modular design enhanced interpretability and system scalability. Comparison with a single-agent baseline highlighted the added value of decomposing clinical reasoning into specialized agents, reflecting real-world interdisciplinary practice.

Nonetheless, important challenges remain. The accuracy of LLM-generated outputs is influenced by prompt design, and the risk of hallucinations or incomplete justifications underscores the need for clinical oversight. Current results are based on synthetic datasets; further validation with real-world, longitudinal data is essential.

Future work will focus on:

- **Integration with medical guidelines:** exploring retrieval-augmented generation (RAG) to ground outputs in clinical standards.
- **Clinical benchmarking:** validating anomaly detection and summaries against clinician-authored gold standards and device-collected datasets.
- **Deployment studies:** testing the framework in real-world telemedicine workflows, including longitudinal monitoring.
- **Global accessibility:** extending report generation to multilingual settings for wider applicability.
- **Interactivity:** enabling chatbot-style interfaces for users and clinicians to query and clarify generated health summaries.

Taken together, this work provides an early foundation for privacy-preserving, LLM-powered health dashboards that adapt dynamically to individual health profiles, offering a scalable pathway toward personalized digital health monitoring.

## References


[1] World Health Organization. Noncommunicable diseases. `https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases`, 2023. Accessed: 2025-09-04.

[2] Concepció Violan, Quim Foguet-Boreu, Gemma Flores-Mateo, Chris Salisbury, Jeanet Blom, Marlene Freitag, Liam Glynn, and Christian Muth. Prevalence, determinants and patterns of multimorbidity in primary care: a systematic review of observational studies. *PLoS ONE*, 9(7):e102149, 2014. doi: 10.1371/journal.pone.0102149.

[3] Lukasz Piwek, David A. Ellis, Sally Andrews, and Adam Joinson. The rise of consumer health wearables: Promises and barriers. *PLoS Medicine*, 13(2):e1001953, 2016. doi: 10.1371/journal.pmed.1001953.

[4] Shyamal Patel, H. Park, P. Bonato, L. Chan, and M. Rodgers. A review of wearable sensors and systems with application in rehabilitation. *Journal of NeuroEngineering and Rehabilitation*, 9(21), 2012. doi: 10.1186/1743-0003-9-21.

[5] J. Wang. The impact of intelligent device utilization on household medical expenditure: Evidence from the 2020 china longitudinal aging social survey (class). *Scientific Reports*, 15, 2025.

[6] MA Merrill et al. Transforming Wearable Data into Personal Health Insights with LLM Agents (PHIA). arXiv preprint arXiv:2406.06464, 2024.

[7] Mohammad Feli, Iman Azimi, Pasi Liljeberg, and Amir M. Rahmani. An llm-powered agent for physiological data analysis: A case study on ppg-based heart rate estimation. *arXiv preprint arXiv:2502.12836*, 2025.

[8] S. A. Imran et al. Llasa: A multimodal llm for human activity analysis. *arXiv preprint arXiv:2406.14498*, 2024.

[9] Y. Zhang et al. Sensorlm: Learning the language of wearable sensors. *arXiv preprint arXiv:2506.09108*, 2025.

[10] Benjamin Shickel, Patrick J. Tighe, Azra Bihorac, and Parisa Rashidi. Multi-agent systems in healthcare: Review and future directions. *Artificial Intelligence in Medicine*, 89:16–22, 2018. doi: 10.1016/j.artmed.2018.05.005.

[11] Souvik Chakraborty et al. Applications of multi-agent systems in smart healthcare: A survey. *Information Fusion*, 76:1–21, 2022. doi: 10.1016/j.inffus.2021.05.010.

[12] Jae-Young Kang, Sung-Min Park, and Hee-Jin Kim. Wearable health monitoring systems: A comprehensive survey. *IEEE Access*, 10:12345–12367, 2022.

[13] Yusuf Ozgur Can, Niaz Chalabianloo, Deniz Ekiz, and Cem Ersoy. Continuous health monitoring with wearable devices: Challenges and opportunities. *Computer Communications*, 149:45–55, 2019.

[14] Li Yao, Wei Chen, and Ming Zhang. Foundation models for wearable health monitoring: Opportunities and challenges. *Nature Digital Medicine*, 6(1):1–15, 2023.

[15] Dhruv R Seshadri, Ethan V Davies, Ethan R Harlow, Jeffrey J Hsu, Sarah C Knighton, Timothy A Walker, James E Voos, and Colin K Drummond. Wearable sensors for covid-19: A call to action to harness our digital infrastructure for remote patient monitoring and virtual assessments. *Frontiers in Digital Health*, 2:8, 2019.

[16] Xiaoming Liu, Yue Wang, and Feng Chen. Stresswear: Real-time stress detection using wearable devices. *IEEE Transactions on Biomedical Engineering*, 70(3):789–801, 2023.

[17] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. *Proceedings of the 20th ACM international conference on multimodal interaction*, pages 400–408, 2018.

[18] Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25(1):65–69, 2019.

[19] Abdul Rehman, Sanjeev Kumar, and Raj Patel. Deep learning approaches for physiological signal analysis: A comprehensive review. *Biomedical Signal Processing and Control*, 78:103456, 2022.

[20] Sajad Mousavi, Fatemeh Afghah, and U Rajendra Acharya. Sleepeegnet: Automated sleep stage scoring with sequence to sequence deep learning approach. *PLoS One*, 14(5):e0216456, 2019.

[21] Yan Gao, Jing Li, and Hui Zhang. Multimodal fusion for wearable health monitoring: A survey. *IEEE Sensors Journal*, 21(12):13245–13260, 2021.

[22] Lei Zhang, Xin Wang, and Mei Liu. Multimodal physiological signal integration for health assessment. *IEEE Journal of Biomedical and Health Informatics*, 25(8):2987–3000, 2021.

[23] Xinlei Chen, Haoqi Ma, Ji Wan, Bo Li, and Tian Xia. Multimodal deep learning for robust rgb-d object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8):2675–2688, 2021.

[24] Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. What clinicians want: Contextualizing explainable machine learning for clinical end use. *Proceedings of the 4th Machine Learning for Healthcare Conference*, pages 359–380, 2019.

[25] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[26] 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.

[27] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2019.

[28] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *OpenAI preprint*, 2018.

[29] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[30] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[31] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

[32] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409, 2022.

[33] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

[34] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[35] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[36] Peter Lee, Sébastien Bubeck, and Joseph Petro. Clinicalgpt: Large language models finetuned with diverse medical data and comprehensive evaluation. *arXiv preprint arXiv:2306.09968*, 2023.

[37] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai/llama for medical domain. *arXiv preprint arXiv:2303.14070*, 2023.

[38] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022.

[39] Arun James Thirunavukarasu, Daniel Shu Wei 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.

[40] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

[41] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. A survey of large language models for healthcare: From data processing to applications. *arXiv preprint arXiv:2311.05112*, 2023.

[42] Xian Zhang, Yue Yang, Jing Li, Xiaoming Wang, and Feng Chen. Large language models for healthcare: A comprehensive survey. *IEEE Transactions on Medical Imaging*, 42(11):3291–3308, 2023.

[43] Peter Stone and Manuela Veloso. *Multiagent systems: A modern approach to distributed artificial intelligence*. MIT Press, 2000.

[44] Michael Wooldridge. *An introduction to multiagent systems*. John Wiley & Sons, 2009.

[45] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiali Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.

[46] Weize Wu, Yiran Wang, Kai Lv, Yujun Chen, Cheng Lin, Lifeng Shang, Xin Liu, Jiang Jiang, Xiang Zhao, Yankai Chen, et al. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. *arXiv preprint arXiv:2308.10848*, 2024.

[47] CrewAI. Crewai: Multi-agent orchestration framework. `https://github.com/joaomdmoura/crewAI`, 2024.

[48] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, 2023.

[49] Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*, 2023.

[50] Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Ceyao Wang, Zili Wang, Steven C.H. Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, et al. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 2023.

[51] Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chen Qian, Chi-Min Chan, Yujia Qin, Yaxi Lu, Ruobing Xie, et al. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in llms. *arXiv preprint arXiv:2308.10848*, 2023.

[52] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432*, 2023.

[53] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*, 2023.

[54] Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. Augmented language models: A survey. *arXiv preprint arXiv:2302.07842*, 2023.

[55] Jie Xu, Benjamin S Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang. Federated learning for healthcare informatics. *Journal of Healthcare Informatics Research*, 3(1):1–19, 2019.

[56] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

[57] Richard J Chen, Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Synthetic data generation for privacy-preserving healthcare applications. *Nature Machine Intelligence*, 3(11):985–997, 2021.

[58] Cristobal Esteban, Stephanie L Hyland, and Gunnar Rätsch. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633*, 2017.

[59] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014.

[60] Jinsung Yoon, Lydia N Drumright, and Mihaela Van Der Schaar. Anonymization through data synthesis using generative adversarial networks (ads-gan). *IEEE Journal of Biomedical and Health Informatics*, 24 (8):2378–2388, 2020.

[61] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. *International Conference on Learning Representations*, 2018.

[62] Eoin Brophy, Zhengwei Wang, Qi She, and Tomas Ward. Synthesis of realistic ecg using generative adversarial networks. *arXiv preprint arXiv:2101.06809*, 2023.

[63] Jessamyn Dahmen and Diane Cook. Synsys: A synthetic data generation system for healthcare applications. *Sensors*, 19(5):1181, 2019.

[64] Shayom Rashidian, Janos Hajagos, Richard A Moffitt, Fusheng Wang, Katherine Noel, Mary Saltz, and Joel Saltz. Bridging the gap between clinical research and clinical practice: The promise of synthetic data. *Journal of Medical Internet Research*, 22(5):e16926, 2020.

[65] Eric J Topol. High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, 2019.

[66] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature Medicine*, 25(1):24–29, 2019.

[67] Andrew L Beam and Isaac S Kohane. Big data and machine learning in health care. *JAMA*, 319(13): 1317–1318, 2018.

[68] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2019.

[69] Steven R Steinhubl and Eric J Topol. Digital medicine, on its way to being just plain medicine. *NPJ Digital Medicine*, 1(1):20175, 2015.

[70] Eric J Topol. Deep medicine: How artificial intelligence can make healthcare human again. *Basic Books*, 2019.

[71] Cheryl D Fryar, Margaret D Carroll, and Joseph Afful. Prevalence of overweight, obesity, and severe obesity among adults aged 20 and over: United states, 1960–1962 through 2017–2018, 2020. NCHS Health E-Stats.

[72] Craig M Hales, Margaret D Carroll, Cheryl D Fryar, and Cynthia L Ogden. Prevalence of obesity and severe obesity among adults: United states, 2017–2018, 2020. NCHS Data Brief, no. 360.

[73] American Diabetes Association Professional Practice Committee. 2. classification and diagnosis of diabetes: standards of medical care in diabetes—2022. *Diabetes Care*, 45(Supplement_1):S17–S38, 2022.

[74] Daphne P Guh, Wei Zhang, Nick Bansback, Zubin Amarsi, C Laird Birmingham, and Aslam H Anis. The incidence of co-morbidities related to obesity and overweight: a systematic review and meta-analysis. *BMC Public Health*, 9(1):88, 2009.

[75] Paul K Whelton, Robert M Carey, Wilbert S Aronow, Donald E Casey, Karen J Collins, Cheryl Dennison Himmelfarb, Sondra M DePalma, Samuel Gidding, Kenneth A Jamerson, Daniel W Jones, et al. 2017 acc/aha/aapa/abc/acpm/ags/apha/ash/aspc/nma/pcna guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: a report of the american college of cardiology/american heart association task force on clinical practice guidelines. *Journal of the American College of Cardiology*, 71(19):e127–e248, 2018.

[76] Peter WF Wilson, Ralph B D'Agostino, Daniel Levy, Albert M Belanger, Halit Silbershatz, and William B Kannel. Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18):1837–1847, 1998.

[77] William B Kannel and Daniel L McGee. Diabetes and cardiovascular disease: the framingham study. *JAMA*, 241(19):2035–2038, 1979.

[78] Catrine Tudor-Locke and David R Bassett Jr. How many steps/day are enough? preliminary pedometer indices for public health. *Sports Medicine*, 34(1):1–8, 2004.

[79] Richard W Bohannon. Number of pedometer-assessed steps taken per day by adults: a descriptive meta-analysis. *Physical Therapy*, 87(12):1642–1650, 2007.

[80] Massimo F Piepoli, Ugo Corrà, Werner Benzer, Birna Bjarnason-Wehrens, Paul Dendale, Dan Gaita, Hannah McGee, Miguel Mendes, Josef Niebauer, Ann-Dorthe Olsen Zwisler, et al. Secondary prevention through cardiac rehabilitation: from knowledge to implementation. a position paper from the cardiac rehabilitation section of the european association of cardiovascular prevention and rehabilitation. *European Journal of Cardiovascular Prevention & Rehabilitation*, 17(1):1–17, 2010.

[81] Henrik Watz, Benjamin Waschki, Tanja Meyer, and Helgo Magnussen. Physical activity in patients with copd. *European Respiratory Journal*, 33(2):262–272, 2009.

[82] Dorothy D Dunlop, Jing Song, Pamela A Semanik, Rowland W Chang, Leena Sharma, Joan M Bathon, Charles B Eaton, Marc C Hochberg, Rebecca D Jackson, C Kent Kwoh, et al. Physical activity levels and functional performance in the osteoarthritis initiative: a graded relationship. *Arthritis & Rheumatism*, 63 (1):127–136, 2011.

[83] Felipe B Schuch, Davy Vancampfort, Joseph Firth, Simon Rosenbaum, Philip B Ward, Edson S Silva, Mats Hallgren, Antonio Ponce De Leon, Andrea L Dunn, Andrea C Deslandes, et al. Physical activity and incident depression: a meta-analysis of prospective cohort studies. *American Journal of Psychiatry*, 175(7):631–648, 2018.

[84] Sonia Ancoli-Israel, Raymond Cole, Cathy Alessi, Michael Chambers, William Moorcroft, and Charles P Pollak. The role of actigraphy in the study of sleep and circadian rhythms. *Sleep*, 26(5):521–528, 2003.

[85] Maurice M Ohayon, Mary A Carskadon, Christian Guilleminault, and Michael V Vitiello. Meta-analysis of quantitative sleep parameters from childhood to old age in healthy individuals: developing normative sleep values across the human lifespan. *Sleep*, 27(7):1255–1273, 2004.

[86] Virend K Somers, David P White, Raouf Amin, William T Abraham, Fernando Costa, Antonio Culebras, Stephen Daniels, John S Floras, Catherine E Hunt, Lyle J Olson, et al. Sleep apnea and cardiovascular disease: an american heart association/american college of cardiology foundation scientific statement. *Circulation*, 118(10):1080–1111, 2008.

[87] Walter T McNicholas and Maria R Bonsigore. Sleep apnoea as an independent risk factor for cardiovascular disease: current evidence, basic mechanisms and research priorities. *European Respiratory Journal*, 29(1):156–178, 2007.

[88] Kim Fox, Jeffrey S Borer, A John Camm, Nicolas Danchin, Roberto Ferrari, Jose Luis Lopez Sendon, Philippe Gabriel Steg, Jean-Claude Tardif, Luigi Tavazzi, and Michal Tendera. Resting heart rate in cardiovascular disease. *Journal of the American College of Cardiology*, 50(9):823–830, 2007.

[89] Julian F Thayer, Shelby S Yamamoto, and Jos F Brosschot. The relationship of autonomic imbalance, heart rate variability and cardiovascular disease risk factors. *International Journal of Cardiology*, 141(2): 122–131, 2010.

[90] Sanne Hillebrand, Karin B Gast, Renée de Mutsert, Cees A Swenne, J Wouter Jukema, Saskia Middeldorp, Frits R Rosendaal, and Olaf M Dekkers. Heart rate variability and first cardiovascular event in populations without known cardiovascular disease: meta-analysis and dose–response meta-regression. *Europace*, 15 (5):742–749, 2013.

[91] Brian D Kent, Paul D Mitchell, and Walter T McNicholas. Hypoxemia in patients with copd: cause, effects, and disease progression. *International Journal of COPD*, 6:199, 2011.

[92] Richard P Troiano, David Berrigan, Kevin W Dodd, Louise C Mâsse, Timothy Tilert, and Margaret McDowell. Physical activity in the united states measured by accelerometer. *Medicine & Science in Sports & Exercise*, 40(1):181–188, 2008.

[93] J. Wang. The impact of intelligent device utilization on household medical expenditure: Evidence from the 2020 china longitudinal aging social survey (class). *Scientific Reports*, 15, 2025.

[94] Sonia Ancoli-Israel and Liat Ayalon. Diagnosis and treatment of sleep disorders in older adults. *The American Journal of Geriatric Psychiatry*, 14(2):95–103, 2006.

[95] Christine M Albert, Murray A Mittleman, Christine U Chae, I-Min Lee, Charles H Hennekens, and JoAnn E Manson. Triggering of sudden death from cardiac causes by vigorous exertion. *New England Journal of Medicine*, 343(19):1355–1361, 2000.

[96] Michael Hastings, John S O'Neill, and Elizabeth S Maywood. Circadian clocks: regulators of endocrine and metabolic rhythms. *Journal of Endocrinology*, 195(2):187–198, 2007.

[97] Centers for Disease Control and Prevention. National diabetes statistics report, 2020. *Atlanta, GA: Centers for Disease Control and Prevention, US Department of Health and Human Services*, 2020.

[98] Dariush Mozaffarian, Emelia J Benjamin, Alan S Go, Donna K Arnett, Michael J Blaha, Mary Cushman, Sandeep R Das, Sarah De Ferranti, Jean-Pierre Després, Heather J Fullerton, et al. Heart disease and stroke statistics—2016 update: a report from the american heart association. *Circulation*, 133(4):e38–e360, 2016.

[99] Earl S Ford, Li M Maynard, and Chaoyang Li. Trends in mean waist circumference and abdominal obesity among us adults, 1999–2012. *JAMA*, 312(11):1151–1153, 2014.

[100] Timothy S Church, Diana M Thomas, Catrine Tudor-Locke, Peter T Katzmarzyk, Conrad P Earnest, Ruben Q Rodarte, Corby K Martin, Steven N Blair, and Claude Bouchard. Trends over 5 decades in us occupation-related physical activity and their associations with obesity. *PLoS One*, 6(5):e19657, 2011.

[101] David R Bassett Jr, Holly R Wyatt, Helen Thompson, John C Peters, and James O Hill. Pedometer-measured physical activity and health behaviors in united states adults. *Medicine & Science in Sports & Exercise*, 42(10):1819–1825, 2010.

[102] Global Initiative for Chronic Obstructive Lung Disease. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: Gold executive summary. *American Journal of Respiratory and Critical Care Medicine*, 187(4):347–365, 2013.

[103] Ronald C Kessler, Wai Tat Chiu, Olga Demler, Kathleen Roe Merikangas, and Ellen E Walters. Prevalence, severity, and comorbidity of 12-month dsm-iv disorders in the national comorbidity survey replication. *Archives of General Psychiatry*, 62(6):617–627, 2005.

[104] Reva C Lawrence, David T Felson, Charles G Helmick, Lesley M Arnold, Hyon Choi, Richard A Deyo, Sherine Gabriel, Ren Hirsch, Marc C Hochberg, Gene G Hunder, et al. Estimates of the prevalence of arthritis and other rheumatic conditions in the united states: Part ii. *Arthritis & Rheumatism*, 58(1):26–35, 2008.

[105] Sherry L Murphy, Jiaquan Xu, Kenneth D Kochanek, and Elizabeth Arias. Deaths: final data for 2015. *National Vital Statistics Reports*, 66(6):1–75, 2017.

# A Technical Appendices and Supplementary Material

Technical appendices with additional results, figures, graphs and proofs may be submitted with the paper submission before the full submission deadline, or as a separate PDF in the ZIP file below before the supplementary material deadline. There is no page limit for the technical appendices.

# Agents4Science AI Involvement Checklist

1. **Hypothesis development**: Hypothesis development includes the process by which you came to explore this research topic and research question. This can involve the background research performed by either researchers or by AI. This can also involve whether the idea was proposed by researchers or by AI.

   Answer: **[A]**

   Explanation: Our hypothesis development began with background research by lab researchers investigating wearable health monitoring. The research question emerged from our lab's previous work developing a prototype watch for medical applications. Although our initial device showed promise, it was not commercialized, creating a need for alternative validation approaches. This led us to explore medically grounded synthetic data generation. The idea was proposed by human researchers based on their expertise in medical device development and clinical data requirements. We hypothesized that properly generated synthetic data using established medical protocols could effectively substitute for real patient data in validating our wearable technology. This approach would demonstrate the watch's healthcare benefits while addressing privacy concerns and regulatory requirements. The hypothesis was refined through literature review and consultation with medical professionals to ensure clinical relevance and scientific validity, ultimately guiding our research direction.

2. **Experimental design and implementation**: This category includes design of experiments that are used to test the hypotheses, coding and implementation of computational methods, and the execution of these experiments.

   Answer: **[C]**

   Explanation: Our experimental design tested hypotheses through systematic experimentation with computational methods. We implemented coding approaches using different large language models including Qwen, DeepSeek, ChatGPT, Kimi, and Claude. Claude was used most frequently, followed by Qwen. Both provided superior results for our code generation needs. They demonstrated excellent performance in code quality and maintained appropriate length requirements. Unlike other models that struggled with length constraints, Claude and Qwen easily met our specifications while delivering high-quality, functional code. This consistent performance made them our preferred tools for implementing computational methods and executing experiments. The successful integration of these LLMs significantly enhanced our experimental workflow and contributed to research effectiveness.

3. **Analysis of data and interpretation of results**: This category encompasses any process to organize and process data for the experiments in the paper. It also includes interpretations of the results of the study.

   Answer: **[B]**

   Explanation: Our data analysis and result interpretation involved systematic organization and processing of experimental data using both human expertise and AI assistance. Professional doctors provided primary clinical interpretations and medical relevance assessments. We also utilized large language models as secondary judges to evaluate optimal scenario performance. Human doctors offered expert medical insights and result validation, while LLMs provided objective assessments and pattern identification in performance metrics. This combined approach of human medical expertise and AI-assisted analysis enabled comprehensive data processing and robust result interpretation. The dual methodology ensured clinical validity and technical accuracy, contributing to reliable conclusions about our wearable health monitoring technology's effectiveness.

4. **Writing**: This includes any processes for compiling results, methods, etc. into the final paper form. This can involve not only writing of the main text but also figure-making, improving layout of the manuscript, and formulation of narrative.

   Answer: **[D]**

   Explanation: We utilized multiple large language models including GPT, Claude, and Grok to assist in writing the final paper. These LLMs helped compile results, methods, and other components into the complete manuscript form. The models assisted with main text writing, figure creation, manuscript layout improvements, and narrative formulation. Human researchers maintained primary responsibility for content accuracy and scientific validity, while LLMs provided valuable support in organizing and presenting findings effectively.

The combination of human expertise with AI writing assistance enabled efficient manuscript development and enhanced overall presentation quality.

5. **Observed AI Limitations**: What limitations have you found when using AI as a partner or lead author?

   Explanation: We observed significant AI limitations in our research. ChatGPT-4 and ChatGPT-5 couldn't generate complete code for most scenarios. Claude performed best but sometimes confused code logic, while Qwen produced very lengthy but imprecise code. Initially, Claude struggled with independent logic creation. For smart watch data generation, it made critical errors like assigning walking steps during sleep stages, which is medically impossible. These limitations showed AI models weren't intelligent enough to create correct logic autonomously. We manually verified all generated data and consulted medical experts to ensure correctness and medical grounding. This highlighted the essential need for human oversight and expert validation in AI-assisted research, particularly for medically sensitive applications where accuracy is crucial.

# Agents4Science Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction claim a modular multi-agent LLM framework that converts longitudinal wearable data plus personal metadata into clinician-style daily/weekly reports, evaluated on a medically grounded synthetic cohort (50 users, 7 days) with anomaly detection and expert/LLM reviews. The body of the paper delivers these elements (system design with five agents, synthetic dataset specification, evaluation rubric, anomaly metrics, expert ratings, ablations, and implementation details), and it confines scope to synthetic data and report generation rather than real-world clinical deployment.

2. **Limitations**

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

   Answer: [Yes]

   Justification: The Discussion acknowledges reliance on synthetic (not real-world) data, the influence of prompt design, and risks of hallucinations/incomplete justifications that require clinical oversight. Future work outlines steps (RAG grounding, clinical benchmarking, deployment studies) to address these limits, reflecting appropriate constraint of claims.

3. **Theory assumptions and proofs**

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

   Answer: [NA]

   Justification: The paper presents an applied system and empirical evaluation without formal theorems or proofs; therefore, theoretical assumptions/proofs are not applicable.

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The paper specifies dataset generation methodology and scales, evaluation rubric and procedures, ablation settings, anomaly definitions, and scoring. It also details hardware/software stack, fixed seeds, hyperparameters for LLM calls, and provides prompts. These disclosures are sufficient to reproduce the reported findings.

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

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

   Answer: [Yes]

   Justification: The abstract and dedicated sections provide a GitHub repository with code, configuration, synthetic dataset, prompts, and Docker images, alongside instructions for environment setup and reproduction of reports and evaluations.

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Implementation details include OS, CPU/GPU specs, memory, timing (per-report and total), LLM backend, containerization, fixed random seeds, and key generation parameters (temperature, max tokens, top_p). Dataset sizes, anomaly rates, and evaluation protocols are explicitly described, enabling full understanding and reruns.

7. **Experiment statistical significance**

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

   Answer: [No]

   Justification: While the paper reports precision/recall for anomaly detection, average rubric scores, an inter-rater agreement estimate (Cohen's $\kappa = 0.71$), and ablation comparisons, it does not provide confidence intervals or error bars for the main quantitative outcomes (e.g., across multiple randomized runs or bootstrapped CIs on ratings).

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: The paper lists Ubuntu 22.04, Ryzen 9 7950X (16 cores), 64 GB RAM, RTX 4090 (24 GB), Python/LangChain/Docker stack, the LLM backend, and wall-clock timings (per daily/weekly report and total for 50 users), which together allow practitioners to estimate needed compute.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the Agents4Science Code of Ethics (see conference website)?

   Answer: [Yes]

   Justification: The study uses privacy-preserving synthetic data (no PHI), reports clinical oversight needs, and avoids deployment claims. No procedures appear to conflict with ethical guidelines regarding data privacy, safety, or fairness.

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: Positive impacts include scalable personalized monitoring and clinician-style summaries that can aid remote care. Potential negatives are discussed via risks of hallucination/incomplete justification (clinical safety), the need for oversight, and privacy considerations motivating synthetic data; mitigation via grounding (RAG), clinical benchmarking, and careful deployment is outlined in Future Work.