Imperial College London
Department of Brain Sciences

Developing a Foundation Model in In-home Monitoring Data for Healthcare Applications

MPhil Thesis

Jin Cui

Supervisors:

Prof. Payam Barnaghi
Prof. Gregory Scott

January 26, 2026

Statement of Originality

I hereby declare that this thesis is my own work and contains no material previously published or written by another person, except where due acknowledgement has been made. The content of this thesis is the result of work I have carried out since the commencement of my research degree and has not been submitted for any other degree or professional qualification.

Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Abstract

We propose a foundation model for in-home monitoring data to identify behavioral patterns and verify their clinical relevance in people living with dementia (PLWD). This method combines self-supervised representation learning with clinical outcome-oriented validation, focusing on agitation, urinary tract infections (UTIs), and cognitive decline. We encode multimodal patient data (sensors, activity patterns, EHRs) into structured latent representations using a pre-trained language model and a PageRank-based state transition model. Using Retrieval Augmented Generation (RAG), we synthesize privacy-preserving virtual cohorts statistically consistent with real data.

Validation indicates that this model outperforms traditional methods in predicting agitation, UTIs, and cognitive scores. Specifically, the ADAS-Cog prediction MAE was reduced to 10.46 (95% CI: 8.31-12.63), compared to 11.41 for the baseline. For UTI prediction, the model achieved 0.795 accuracy on real data and 0.906 on synthetic data. Sensitivity and PPV on synthetic data reached 0.909 and 0.957, respectively, surpassing real data metrics. The generated virtual cohort demonstrated high fidelity, achieving a Jensen-Shannon Distance of 0.2623 and Frechet Inception Distance (FID) of 0.0212. While agitation classification remained challenging (accuracy 52.3%), highlighting the need for prior knowledge, the overall two-stage coding architecture successfully extracted deep behavioral patterns. This work advances personalized dementia care by providing clinically instructive insights and secure, large-scale health monitoring through high-fidelity virtual cohort generation.

Publications

- Cui, J., Capstick, A., Barnaghi, P., Scott, G.. Two-Stage Representation Learning for Analyzing Movement Behavior Dynamics in People Living with Dementia. AAAI 2025 workshop on Large Language Models and Generative AI for Health. https://openreview.net/forum?id=wBpci9Zq0i

In preparation

- Cui, J., Capstick, A., Barnaghi, P., Scott, G.. Post-Training Retrieval-Augmented Generation: A Method for Building Structured Virtual Cohorts to Achieve Healthcare Temporal Data Anonymization and Augmentation.

List of Figures

1 An illustration of in-home monitoring to analyze behavioral patterns of People Living with Dementia. The graph is collected from `https://www.ukdri.ac.uk/centres/care-research-technology/minder` 1
2 Minder sensors 8
3 Daily histogram 9
4 Location histogram showing the distribution of 24,467,307 individual records across different semantic zones. 9
5 Top 50 timeseries distribution 10
6 Timeseries of cognitive test of the test set participants 11
7 Cohort Demographics 11
8 Monthly Distribution of Records 12
9 Encoding model pipeline 13
10 Pretrained BERT model training strategy 14
11 Data Processing and Synthetic Data Generation Visualization 17
12 Visualization of the PageRank computation process. Panel (a) illustrates the patient's movement trajectory projected into a 2D t-SNE latent space, where clusters represent distinct behavioral patterns. Panel (b) visualizes the computed PageRank values, where the size of each node corresponds to the stationary probability (importance) of that state in the patient's daily routine. 19
13 Encoding Model Pipeline 21
14 Decoding Model Pipeline 21
15 Unsupervised Methods 23
16 Downstream Tasks Pipeline 24
17 T-SNE for embedded daily movement strings in the test set 25
18 Multi-period participant deep state vector and similarity 25
19 PCA Visualization of Real vs. Generated PLWD Movement Trajectory Embeddings (all-MiniLM-L12-v2) 28
20 Model Comparison Heatmap 29
21 ADAS-Cog Prognostic Results on Test set 32
22 MMSE Prognostic Results on Test set 33
23 NPI Prognostic Results on Test set 34
24 Bland–Altman for UTI 35
25 UTI classification results on the test set, comparing real and virtual cohort predictions across multiple patients. 63
26 Bland–Altman for Agitation 64
27 Agitation classification results on the test set, comparing real and virtual cohort predictions across multiple patients. 65
28 MAE Comparison of Real vs. Virtual Cohorts(Largest & Smallest Diff) . 66
29 Location timeseries for articipants in cluster 1 66
30 Location timeseries for participants in cluster 2 67
31 Location timeseries for participants in cluster 3 67
32 Location timeseries for participants in cluster 4 68

33 Location timeseries for participants in cluster 5 68
34 Mean of location count in five clusters 69
35 Variance of location count in five clusters 70
36 Mean of wake count in five clusters 70
37 T-SNE for individuals in test set 71
38 T-SNE for individuals in test set 72
39 Prediction error for cognitive change (ΔADAS-Cog and ΔMMSE) without current cognitive score in feature set. This figure illustrates the impact on prediction errors (MAE and RMSE) for cognitive change when excluding the current cognitive score from the feature set across different time durations. 73
40 Prediction error for cognitive change (ΔADAS-Cog and ΔMMSE). This figure shows the MAE and RMSE for cognitive change predictions using varying time-series lengths, including baseline, characteristics, and state feature sets. 73
41 Prediction error for current cognitive scores (ADAS-Cog and MMSE). This figure displays the MAE and RMSE for absolute cognitive score predictions, illustrating the performance stability across various time-series lengths with baseline, characteristic, and state feature sets. 74
42 Comparison of calibration plots for ADAS-COG task 74
43 Comparison of calibration plots for MMSE task 75
44 Comparison of calibration plots for NPI task 75
45 Performance comparison across different generative models. The results show that the selection of the generative model has limited impact on key performance trends. 75
46 Clinical Sequence generator example (LLM) 76

List of Tables

1 Overview of the Minder Dataset 6
2 Silhouette scores under different models and parameter settings 15
3 Performance metrics for predicting current ADAS-Cog and MMSE with 95% confidence intervals (CI). The values in parentheses represent the lower and upper bounds of the 95% CI. 27
4 Unsupervised Evaluation Results on Synthetic Dataset 27
5 Model Performance for Different Down Stream Tasks 30
6 Model Performance for Different Down Stream Tasks (Filtered) 30
7 Diagnostic Performance Comparison for UTI Classification 35
8 Intraclass Correlation Coefficients (ICC) for UTI Rater Agreement 36
9 95% Confidence Intervals for ICC Estimates in UTI CLassification 36
10 Diagnostic Performance Comparison for Agitation Classification 37
11 Intraclass Correlation Coefficients (ICC) for Agitation Rater Agreement . 37
12 95% Confidence Intervals for ICC Estimates in Agitation Classification . 38
13 MAE Sensitivity Analysis Results over Different Time Windows (with 95% CI), Grouped by Cohort and Downstream Tasks 40
14 P-values for characteristics comparison with most similar patients. 50
15 P-values for characteristics comparison with least similar patients. 50
16 Effect sizes for comparisons with most similar patients. 50
17 Effect sizes for comparisons with least similar patients. 50
18 Performance metrics for predicting changes of ADAS-Cog and MMSE with 95% confidence intervals (CI). The values in parentheses represent the lower and upper bounds of the 95% CI. The metrics represent the predicted annual change (Δ) for MMSE and ADAS-Cog. 54
19 Performance metrics for predicting changes of ADAS-Cog and MMSE with 95% confidence intervals (CI). The values in parentheses represent the lower and upper bounds of the 95% CI. The metrics represent the predicted annual change (Δ) for MMSE and ADASCOG. Note: This Characteristics feature set does not include the current values of MMSE and ADAS-Cog. 55
20 Performance metrics for predicting current ADAS-Cog and MMSE with 95% confidence intervals (CI). The values in parentheses represent the lower and upper bounds of the 95% CI. 56

Abbreviations

- ADAS-Cog - Alzheimer's Disease Assessment Scale - Cognitive Subscale
- NPI - Neuropsychiatric Inventory
- PLWD - People Living with Dementia
- MMSE - Mini-Mental State Examination
- UTI - Urinary Tract Infection
- EHR - Electronic Health Records
- LLM - Large Language Model
- RAG - Retrieval-Augmented Generation
- PCA - Principal Component Analysis
- JS - Jensen-Shannon (Distance)
- FID - Frechet Inception Distance
- PIR - Passive Infrared (Sensor)
- HIPAA - Health Insurance Portability and Accountability Act
- GDPR - General Data Protection Regulation
- K-NN - K-Nearest Neighbors
- MLP - Multilayer Perceptron
- MAE - Mean Absolute Error
- RMSE - Root Mean Square Error
- ARIMA - Autoregressive Integrated Moving Average
- GAN - Generative Adversarial Network
- VAE - Variational Autoencoder
- LSTM - Long Short-Term Memory (Network)
- SHAP - SHapley Additive exPlanations

Contents

Abstract ii

Publications iii

List of Figures iv

List of Tables vi

Abbreviations vii

1 Introduction 1
- 1.1 Background and context . . . 1
- 1.2 Aims and objectives . . . 1
- 1.3 Methodology . . . 2
- 1.4 Contributions of this research . . . 2
- 1.5 Structure of this thesis . . . 2

2 Background and Related Work 3
- 2.1 Processing on in-home monitoring time-series data . . . 3
  - 2.1.1 Challenges in Time Series Representation Learning . . . 3
  - 2.1.2 Generative Models for Synthetic Healthcare Data . . . 3
  - 2.1.3 Evaluation Framework . . . 4
  - 2.1.4 Time-Series Forecasting Models . . . 4
  - 2.1.5 Large Language Models in Time-Series Analysis . . . 4
  - 2.1.6 PageRank and Its Applications . . . 5
  - 2.1.7 Synthesis and Augmentation of Healthcare Data . . . 5
- 2.2 Study Data . . . 6
  - 2.2.1 Ethics . . . 6

3 Detailed Data Description 8
- 3.1 Data Collection and details . . . 8

4 Model Development and Architecture 10
- 4.1 Detailed Encoding Model Pipeline . . . 10
- 4.2 Detailed Synthetic Dataset Generation Encoding Process . . . 13
  - 4.2.1 Positive and Negative Sample Selection in Pre-training . . . 14
  - 4.2.2 Sample Selection Workflow . . . 14
  - 4.2.3 Positive & Negative Window Selection Process . . . 15
  - 4.2.4 Embedding and Clustering . . . 16
  - 4.2.5 Pre-Training Process . . . 16
- 4.3 PageRank Iteration for a Single Patient . . . 16
  - 4.3.1 Model Definition . . . 16
  - 4.3.2 Transition Matrix Construction . . . 16
  - 4.3.3 PageRank Computation . . . 17
  - 4.3.4 Algorithm Summary . . . 18
  - 4.3.5 Algorithm Visualization . . . 18

5 Methodology for Building A Latent Representation Model for in-Home monitoring Data **20**
5.1 Preprocessing for Raw Data from Sensors . . . 20
5.2 Data Extraction and Engineering . . . 20
5.3 Data Decoding . . . 21
5.3.1 Retrieval Augmented Generation . . . 22
5.4 Downstream Applications and Validation . . . 22

6 Exploratory Analysis and Downstream Model Development **24**
6.1 Encoder Model . . . 24
6.2 Cognitive Status Prediction: Component Ablation Study . . . 26
6.3 Decoder Model . . . 27
6.3.1 Unsupervised Methods Evaluation . . . 27
6.3.2 Downstream Task Performance . . . 30
6.4 Classification Risk Analysis . . . 31
6.4.1 Urinary Tract Infection Risk Analysis . . . 33
6.4.2 Agitation Risk Analysis . . . 36
6.4.3 Comparison Results . . . 38
6.5 Adversarial Evaluation of Synthetic Data Quality . . . 39
6.5.1 Results and Interpretation . . . 39
6.5.2 Limitations regarding Privacy-Preserving Properties . . . 40
6.6 Sensitivity Analysis of Input Time Windows . . . 40

7 Conclusions and Future Work **42**

Bibliography **47**

A Appendix **48**
A.1 Participants Demographics and Clinical Characteristics . . . 48
A.1.1 Demographics . . . 48
A.1.2 Comorbidities . . . 48
A.1.3 In-Home Movement Trajectories . . . 48
A.1.4 Agitation, Mood, and UTI Status . . . 48
A.2 Distribution of Cognitive and Behavioral Assessments . . . 49
A.2.1 Temporal Distribution of Clinical Records . . . 49
A.3 Datasets and Code Availability . . . 49
A.4 Location time series and distribution visualization for each cluster . . . 49
A.5 T-SNE plots for individual participants in test set . . . 49
A.6 Relations between most similar and dissimilar pariticipants . . . 49
A.7 Clinical Significance of Deep Vector States . . . 51
A.7.1 Model Parameters . . . 52
A.8 LSTM Model for Multivariate Time Series Prediction . . . 53
A.8.1 Ablation Study: Impact of Time-Series Length on Model Performance 56
A.8.2 Prediction of Absolute Cognitive Scores (ADASCOG and MMSE) 57
A.8.3 State Feature Performance . . . 57
A.9 Detailed Data Extracting and Synthetic Data Generation Workflow . . . 57
A.9.1 Raw Data Representation . . . 57
A.9.2 Rectification of Time-Series Data . . . 57
A.9.3 Conversion to Structured Text . . . 58

A.9.4 Retrieval-Augmented Generation (RAG) Query and Generation . 58
A.9.5 Synthetic Data Rectification 58
A.10 Base Generation Model Architecture and Parameters 58
A.11 Downstream Task Model Data Preparation 59
A.11.1 Data Preparation 59
A.11.2 Task Participant Profiling 59
A.11.3 Model Architecture 59
A.11.4 Calibration Results on Regression Tasks 60
A.12 Generation Model Ablation Study 60
A.13 Visualization Design for Clinical Trajectory Chat Framework 61
A.13.1 1. Dialogue Components 61
A.13.2 2. Example Usage 61
A.13.3 Clinical Challenges 62

1 Introduction

Dementia is a progressive neurodegenerative condition that affects memory, cognition, and daily functioning. Beyond cognitive decline, people living with dementia (PLWD) often experience changes in behaviour and activity patterns that can significantly impact their quality of life and the burden of care. Monitoring and understanding these behavioural dynamics is therefore essential both for advancing clinical knowledge and for improving support systems for patients and their caregivers.

1.1 Background and context

In recent years, remote monitoring technologies have provided new opportunities to observe the daily lives of PLWD in a continuous and unobtrusive manner. Wearables and Internet of Things (IoT) devices, deployed within home environments, can capture multimodal time-series data at fine temporal granularity. Such data offer valuable insights into sleep–wake cycles, mobility patterns, and overall behavioural routines, but their complexity poses challenges for direct human interpretation.

The Minder study, developed as part of the UK Dementia Research Institute Care Research and Technology Centre, provides a comprehensive platform for in-home monitoring. As illustrated in Figure 1, it integrates sensor data and contextual information to support analysis of the lived experience of PLWD, with the ultimate goal of enabling early interventions and personalised care.

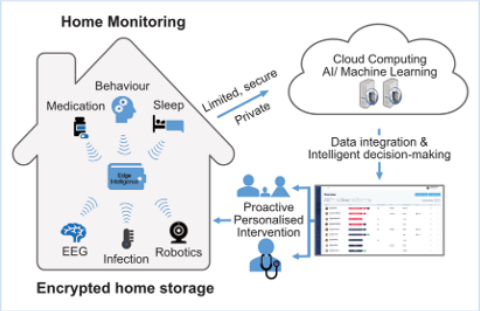


Figure 1: An illustration of in-home monitoring to analyze behavioral patterns of People Living with Dementia. The graph is collected from `https://www.ukdri.ac.uk/centres/care-research-technology/minder`

1.2 Aims and objectives

The aim of this research is to develop data-driven approaches that can capture and interpret the behavioural and cognitive changes of PLWD through home monitoring data. More specifically, this work seeks to:

- Develop representation learning methods for modelling in-home activity and behavioural patterns.
- Validate predictive models for clinically relevant outcomes such as cognitive scores and agitation.

- Explore synthetic cohort generation to support privacy-preserving research and clinical trials.

1.3 Methodology

To address these objectives, this research adopts a multimodal data-driven approach. The study uses continuous time-series data collected from IoT and wearable devices in home settings. Advanced representation learning methods, including language-model-based encoding and graph-based state modelling, are applied to uncover latent structures in behaviour. In addition, privacy-preserving synthetic data generation techniques are explored. The evaluation combines technical validation (e.g., predictive performance, robustness) with clinical validation (e.g., alignment with known outcomes and clinician interpretation).

1.4 Contributions of this research

The research has produced several notable outcomes. A two-stage representation learning framework was developed, which combines pre-trained language models with graph-theoretical analysis to encode and interpret movement trajectories of PLWD. This method improved prediction of cognitive decline and supported the discovery of interpretable behavioural patterns. In addition, a retrieval-augmented generative pipeline was implemented to construct synthetic, clinically consistent cohorts, enabling large-scale analyses while safeguarding privacy. The developed models were validated across multiple clinical tasks, including prediction of urinary tract infections, cognitive scores, and agitation detection, demonstrating both accuracy and practical utility.

1.5 Structure of this thesis

The structure of this thesis is organized as follows. Chapter 2 reviews the background literature and related work on time series data processing, generative models for healthcare, and the application of machine learning in dementia monitoring. Chapter 3 provides a detailed description of the data set used in this study. Chapter 4 explains the development and architecture of the model, including the encoding pipeline and the process for the generation of synthetic data. Chapter 5 details the methodology for building the latent representation model, from pre-processing to downstream validation. Chapter 6 presents the exploratory analysis and results of the downstream model development, including the prediction of cognitive status and the classification risk analysis. Finally, Chapter 7 concludes the thesis by summarizing the main contributions and suggesting directions for future research.

2 Background and Related Work

This research aims to develop a foundation model for analyzing in-home monitoring data of people living with dementia (PLWD), to extract semantically meaningful, privacy-preserving representations that support clinical prediction and behavioral understanding. The model incorporates a large language model (LLM)-based encoding and PageRank-based latent state transition analysis, with additional integration of retrieval-augmented generation (RAG) to construct synthetic virtual cohorts.

Foundation model, which is pre-trained on vast and multi-modality datasets, are difficult to build from scratch, requiring immense computational resources, extensive engineering expertise, and access to massive amounts of data. They are designed to learn general representations that can be applied to downstream tasks, offering a paradigm shift from building task-specific models. While foundation models have seen significant success in deep learning, and are emerging in areas such as genomics, there is a distinct gap in their application to in-home monitoring data for people living with dementia (PLWD), particularly over long time spans. Analyzing such data presents unique challenges due to its temporal complexity, inherent noise, and the critical need for privacy [1]. Our study attempts to build a time-series foundation model for this unique clinical utility, recognizing the potential for personalized insights into dementia progression and behavioral changes. However, a key shortcoming of such foundation models, especially in nascent applications like temporal health data, is their interpretability; while they can identify patterns, fully understanding the "why" behind their predictions or representations remains an active area of research.

2.1 Processing on in-home monitoring time-series data

This section provides an overview of the challenges and related work to process continuous in-home monitoring data for healthcare applicants.

2.1.1 Challenges in Time Series Representation Learning

Training on large-scale, unlabeled time series data presents significant challenges. To address this, semi-supervised and unsupervised approaches are commonly used to encode and extract meaningful features for downstream tasks such as classification and regression. Semi-supervised methods, including nearest neighbor contrastive learning and temporal relation prediction, use labeled and unlabeled data to improve representation quality, improving classification performance [2, 3]. In contrast, unsupervised techniques focus on learning robust features without labels, often utilizing contrastive learning and advanced data augmentation to capture essential temporal patterns [4, 5]. Attention mechanisms and domain-adaptive strategies further refine these representations, improving interpretability and alignment with domain knowledge [6]. However, two key challenges remain: (1) ambiguous labeling criteria in time series data can hinder the effectiveness of the model, and (2) the resulting high-dimensional embeddings remain complex and difficult to interpret [7, 8].

2.1.2 Generative Models for Synthetic Healthcare Data

Rapid advances in machine learning and artificial intelligence (AI) have significantly transformed various domains, including healthcare, where data-driven models play an

increasingly crucial role in clinical decision-making, personalized treatment, and disease prediction [9, 10]. However, effective development of such models depends on the availability of large-scale, high-quality datasets. In healthcare, this is particularly challenging due to the sensitive nature of patient data, strict privacy regulations (e.g. GDPR, HIPAA), and limited data sharing between institutions. These challenges hinder the reproducibility of research and the broader adoption of AI-driven healthcare solutions [11, 12, 13].

To address these limitations, we have turned to generative models to create synthetic datasets that mimic the characteristics of real data while preserving patient privacy. Among these, large language models (LLMs) have emerged as a powerful tool, demonstrating remarkable capabilities to generate high-quality textual and structured data [14]. However, the integration of LLMs with structured and temporal healthcare data remains an underexplored area. Key challenges include ensuring that the generated data not only aligns with real-world distributions but also maintains sufficient diversity, sample quality, and generalization capabilities for downstream applications.

2.1.3 Evaluation Framework

To validate the effectiveness of the proposed framework, we conduct extensive evaluations using both unsupervised metrics and downstream tasks. Unsupervised metrics include sample quality (cosine similarity), diversity (distribution coverage through PCA [15]), generalization (Jensen-Shannon distance [16] and Fréchet Inception Distance [17]), and precision and recall based on k-NN. For downstream tasks, we evaluate the performance of the generated data sets in predicting urinary tract infections (UTI), cognitive evaluations (ADAS-Cog, MMSE, NPI), and behavioral agitation [18]. Our results demonstrate that the generated virtual cohorts not only preserve key statistical properties of real data but also achieve competitive or superior performance in downstream applications.

2.1.4 Time-Series Forecasting Models

Time-series forecasting aims to predict future values based on historical observations. Traditional statistical models, such as the Autoregressive Integrated Moving Average (ARIMA), have been widely used due to their mathematical simplicity and adaptability [19, 20]. Although ARIMA remains effective for data with linear trends, advances in machine learning have introduced more sophisticated models capable of capturing nonlinear dependencies, improving forecast accuracy and robustness [21, 22].

2.1.5 Large Language Models in Time-Series Analysis

The introduction of the Generative Pre-trained Transformer (GPT) by OpenAI marked a transformative shift in natural language processing [23], sparking innovations in large language models (LLMs). Beyond text processing, LLMs are now being explored for time-series analysis. For instance, [24] proposes a framework that reinterprets time-series forecasting as a multipatch prediction task, incorporating a patch-wise decoding layer to enhance temporal sequence learning. Similarly, [25] leverages the autoregressive nature of LLMs for time-series forecasting. In [26], GPT-based text encoding is applied to in-home activity data, enabling vector searching and clustering. Expanding on these approaches, we introduce a secondary modeling stage to enhance interpretability and further analysis.

2.1.6 PageRank and Its Applications

Developed initially to rank web pages, PageRank evaluates the importance of nodes in a directed graph by analyzing link structures [27]. While initially designed for search engines, its applications have extended across various fields. In biological networks, [28] employed personalized PageRank to interpret protein interactions, demonstrating scalable techniques for complex biological data analysis. Similarly, [29] adapted PageRank for metabolic network analysis, illustrating its versatility in modeling intricate systems beyond its original scope.

2.1.7 Synthesis and Augmentation of Healthcare Data

The synthesis and augmentation of healthcare temporal data, particularly for privacy-preserving and machine learning applications, have gained significant attention in recent years.

Anonymization of Temporal Healthcare Data Temporal healthcare data, such as electronic health records (EHRs), physiological signals, and treatment sequences, pose unique challenges to anonymization due to their sequential nature. Various anonymization techniques have been developed to address these challenges. Differential privacy introduces noise to sequences to obscure identifiable features while retaining temporal trends [30]. Clustering-based aggregation groups sequences together to mask unique patterns while preserving analytical utility [31, 32]. In addition, generative adversarial networks (GANs) have been used to perturb data while maintaining temporal fidelity [30]. Despite these advancements, achieving compliance with privacy regulations such as HIPAA and GDPR while ensuring data utility remains an open challenge.

Generative Models for Temporal EHR Data Synthesis Generative models have been widely utilized to synthesize temporal EHR data. GAN-based approaches, such as SynTEG [33], generate realistic temporal sequences while preserving correlations. Variational autoencoder (VAE)-based methods, such as TrialSynth [34], introduce event-level temporal control to enhance data realism. Diffusion models, including EHRPD [35], improve fidelity by modeling predictive denoising processes. Furthermore, hierarchical and autoregressive models, such as HALO [36], generate granular sequences using structured dependencies.

Data Augmentation for Rare Diseases The scarcity of data for rare diseases significantly limits the generalizability of predictive models, prompting the development of data enhancement techniques. Conditional GANs, such as HealthGen [37], generate synthetic underrepresented subpopulations to improve the generalization of the model. Trajectory splitting methods simulate transitions between health states by leveraging subsequence-based approaches [38]. Moreover, multilevel temporal enhancement techniques, such as Multi-TA [39], enhance robustness by applying augmentation at both the event and the trajectory levels.

Retrieval-Augmented Generation (RAG) in Healthcare Retrieval-Augmented Generation (RAG) has recently emerged as a powerful method for healthcare applications, combining pretrained generative models with retrieval mechanisms to ground synthetic

data generation in external knowledge. Biomedical retrieval models such as BiomedRAG integrate disease-specific contexts to support simulation and enhancement tasks. Privacy-preserving frameworks, including SAGE [40], enhance synthetic data generation by using privacy-sensitive retrieval. Furthermore, time-aware embedding techniques condition retrieval results on disease progression, ensuring alignment with clinical hierarchies [41].

Structured Virtual Cohorts in Clinical Applications Structured virtual cohorts have also gained prominence in clinical applications, enabling the emulation of population- or disease-specific distributions to address critical research needs. Synthetic cohorts have been used to mitigate bias by balancing demographic distributions, thus improving the fairness of the model [37]. Virtual cohorts enrich rare disease datasets, enhancing model robustness in underrepresented cases [34]. In addition, clinical validation frameworks such as LS-EHR [42] ensure that synthetic data sets are suitable to validate decision support systems. Despite these advancements, challenges remain in integrating RAG with synthetic cohort construction while ensuring that temporal constraints are maintained during data generation.

2.2 Study Data

The data set used in this study is sourced from the Minder study and contains longitudinal multimodal home monitoring records of PLWD. The table below provides a summary of the characteristics of the dataset.

Table 1: Overview of the Minder Dataset

Property	**Description**
Project Source	Minder Study
Population	134 people living with dementia (PLWD)
Time Range	July 1, 2021 – January 30, 2024
Selected Participants	50 (with complete records)
Total Days Recorded	66,096 days
Sensor Modalities	Passive Infrared (PIR), Door Sensors, Sleep Monitors
Clinical Annotations	MMSE, ADAS-Cog, NPI, UTI, Agitation, HADS
Other Data	Demographics, Comorbidities, Daily Room Transitions
Sampling Granularity	Seconds (up to 86,400 records/day per participant)

2.2.1 Ethics

Remote-monitoring data was collected from households of people living with dementia as part of the ongoing Minder study being conducted at the UK Dementia Research Institute Care Research & Technology centre (see `https://ukdri.ac.uk/centres/care-research-technology` and `https://mindermeetingplace.com`). The London-Surrey Borders Research Ethics Committee formerly South West London Ethics Committee (see `https://www.hra.nhs.uk/about-us/committees-and-services/res-and-recs/search-research-ethics-committees/london-surrey-borders/`) gave ethical approval for the Minder study protocol. The Surrey and Borders NHS Trust information governance committee (for more information go to `https://www.sabp.nhs.uk/aboutus/policies/digital-governance/infoGovPol`) and the Imperial College information

governance team (for more information go to https://www.imperial.ac.uk/admin-services/secretariat/college-governance/charters/policies-regulations-and-codes-of-practice/policy-framework/) reviewed and approved the information governance procedures.

I confirm that all necessary patient/participant consent has been obtained and the appropriate institutional forms have been archived, and that any patient/participant/sample identifiers included were not known to anyone (e.g., hospital staff, patients, or participants themselves) outside the research group so cannot be used to identify individuals.

This section provides the background and reviews related work for developing a foundation model designed to analyze in-home monitoring data from people living with dementia (PLWD). The primary goal is to create semantically meaningful, privacy-preserving representations to support clinical predictions and understand behavior. The text highlights key challenges, including the difficulty of learning from complex and noisy time-series data and the strict privacy regulations (like GDPR and HIPAA) that limit data sharing in healthcare. To overcome these hurdles, the research leverages generative models, specifically Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG), to create synthetic data cohorts. The review covers existing work in time-series analysis, the application of LLMs and PageRank in new domains, various data anonymization and synthesis techniques, and the use of virtual cohorts in clinical research. The chapter concludes by introducing the Minder dataset, the source of the longitudinal monitoring data used in this study. The following chapter will introduce the methodology used in this work.

3 Detailed Data Description

This chapter provide more detailed insights into the data collection process. We then present an explroatory analysis of the data.

3.1 Data Collection and details

The in-home activity monitoring dataset was collected using passive infrared sensors installed in multiple locations within the homes of individuals with dementia, along with sleep pads placed under their mattresses, as illustrated in Figure 2 from the Minder website. These sensors detect motion within a nine-meter range at a maximum angle of 45 degrees diagonally upward. Installed in lounges, kitchens, hallways, bedrooms, and bathrooms, they enable detailed tracking of movement within and between these areas.

We analyzed data recorded from July 1, 2021, to January 30, 2024, covering 66,096 participant-days across 134 individuals. Figure 3 presents the distribution of logged days per participant. Each record includes a participant ID, a timestamp (accurate to the second), the detected activity location, or sleep pad data indicating bed entry and exit events. The dataset comprises 24,467,307 individual records, as shown in Figure 4. Additionally, Figure 5 highlights the top 50 participants with the longest recorded time spans.

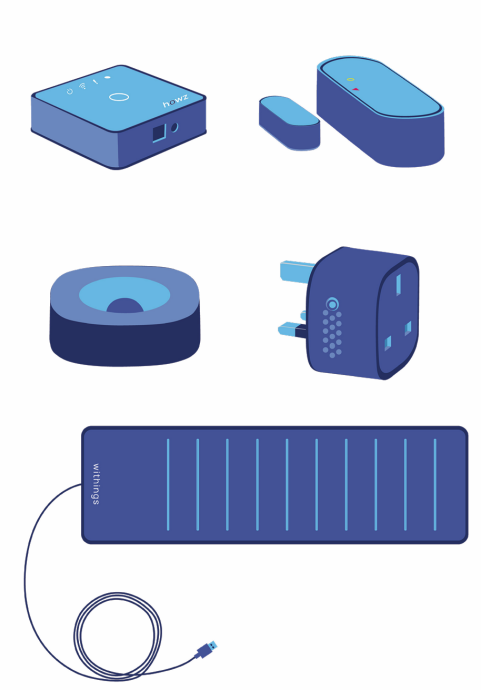


Figure 2: Minder sensors

In addition to activity data, diagnostic information was available for all 134 participants, including birthdate, sex, living situation (e.g., living alone), ethnicity, and diagnosis of dementia. Cognitive assessment scores, such as MMSE and ADAS-Cog, together with their yearly changes, were also included, as shown in Figure 6.

To prevent data leakage, we designated the period from July 1, 2021, to July 1, 2023, for fine-tuning, while data from July 31, 2023, to January 30, 2024, were reserved for testing. The test set was used to extract coding vectors from the fine-tuned language

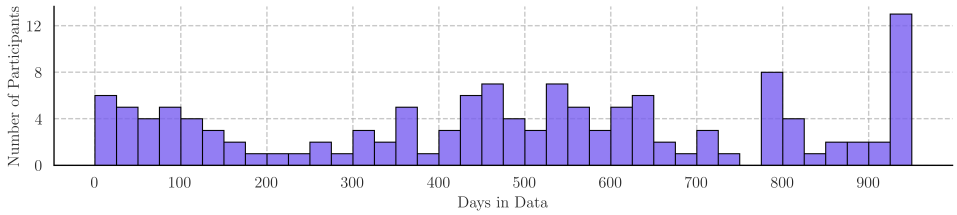


Figure 3: Daily histogram

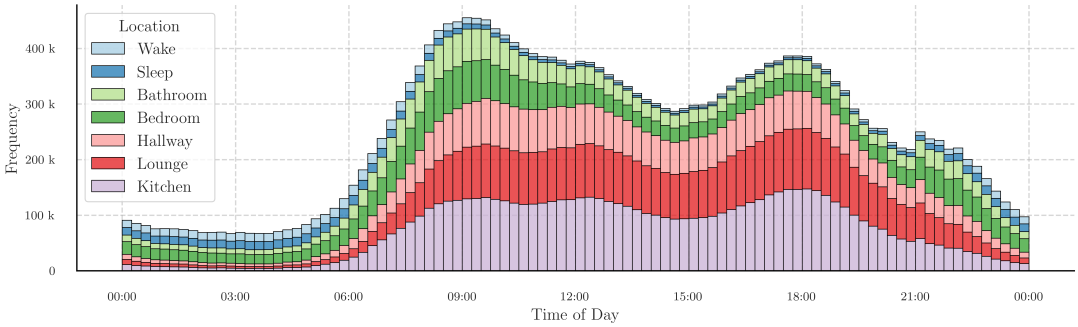


Figure 4: Location histogram showing the distribution of 24,467,307 individual records across different semantic zones.

model and analyze behavioral transfer patterns. Participants with incomplete records, particularly those with data gaps after July 2023, were excluded. The final test set consisted of 50 participants with the most complete post-July 2023 data, totaling 869 comprehensive clinical records.

For our analyses, we used the results of the cognitive tests from August 2023 to January 30, 2024. Furthermore, we incorporated assessments from the previous year to evaluate changes in MMSE and ADAS-Cog scores over time, as shown in Figure 7 and 8.

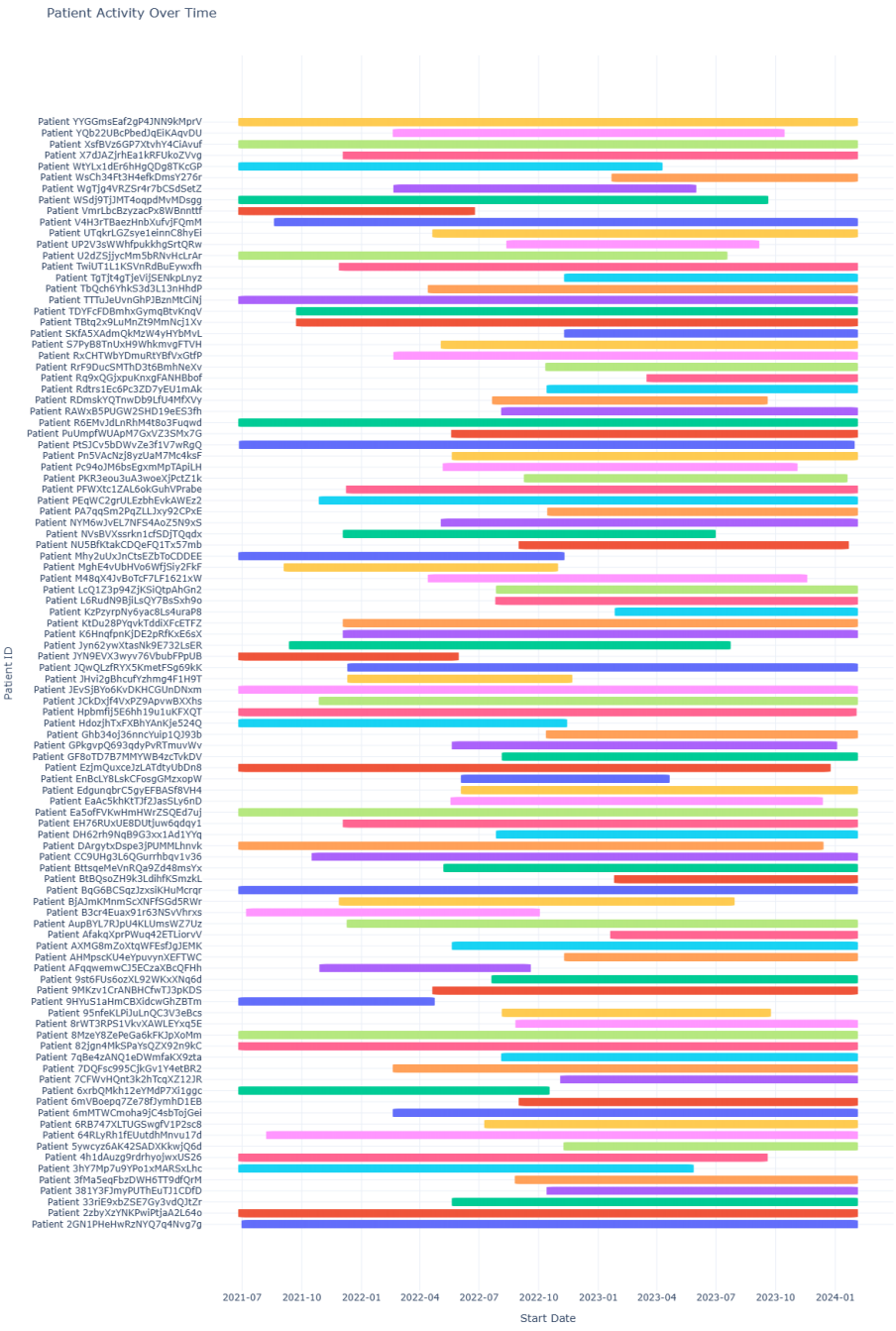


Figure 5: Top 50 timeseries distribution

4 Model Development and Architecture

This chapter explains the basic structure and training strategy of the encoding and decoding model we implemented in this study.

4.1 Detailed Encoding Model Pipeline

As is shown in Figure 9, the workflow for analyzing the dynamics of the PLWD's daily movement behavior involves five primary stages. First, raw temporal data is preprocessed

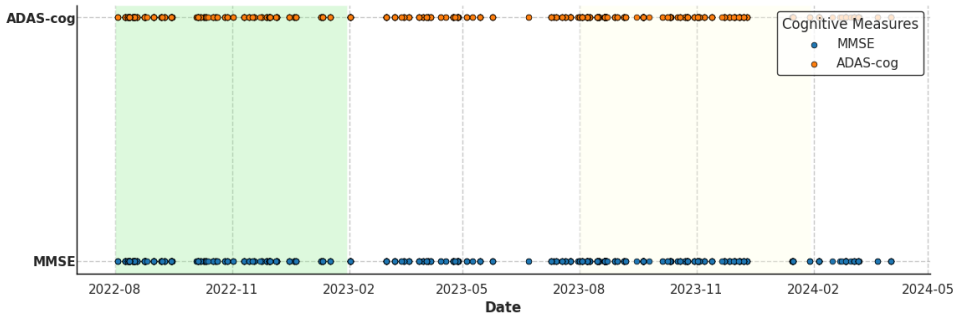


Figure 6: Timeseries of cognitive test of the test set participants

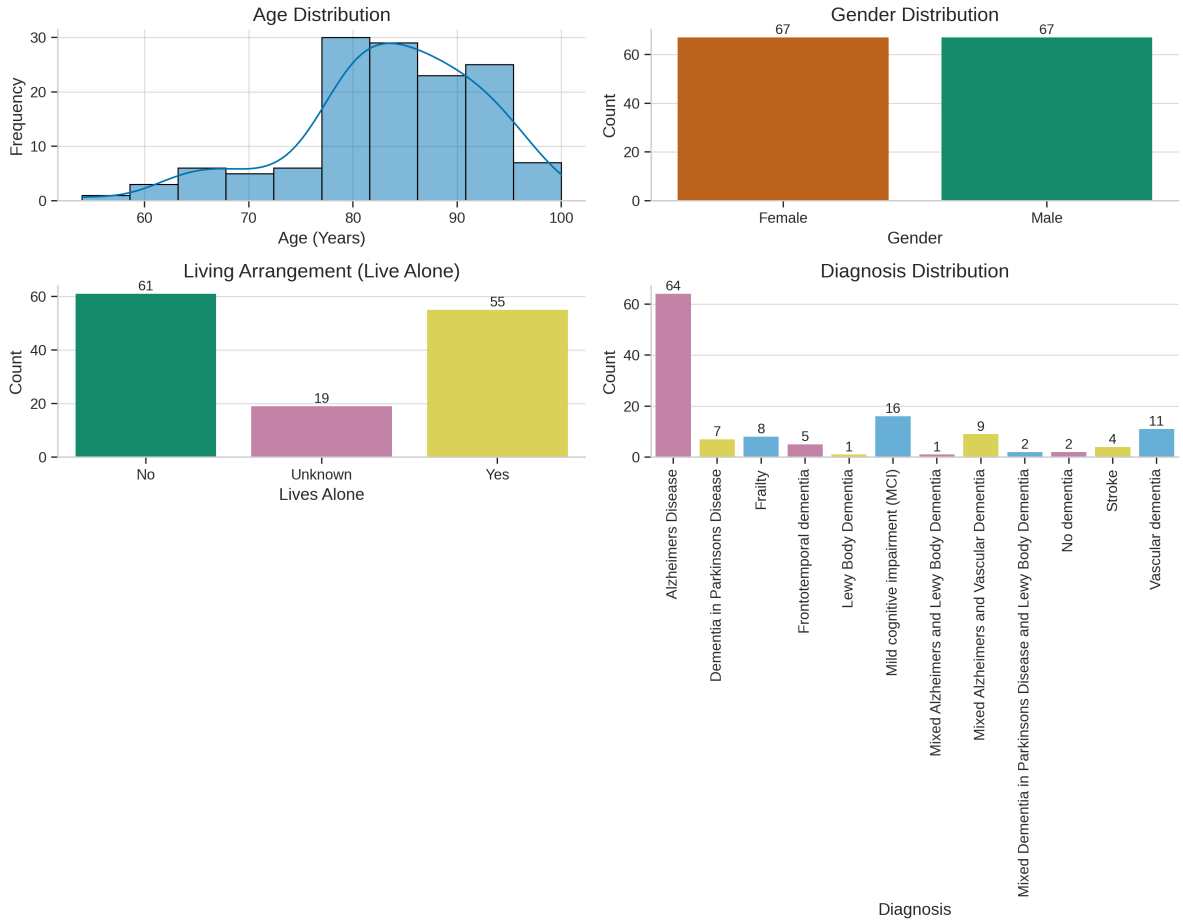


Figure 7: Cohort Demographics

to remove noise and segment multiple-day records into single-day datasets. In the Contrastive Sample Selection phase, one-hot encoding and K-means clustering are applied to create clusters, facilitating the selection of similar and dissimilar samples for contrastive learning. Next, Language Model Encoding utilizes a language model to generate high-dimensional vector representations, with pseudolabeling enhancing the capture of temporal dependencies. The Dimensionality Reduction and Clustering stage uses t-SNE to project embeddings into a 2D space, allowing clustering algorithms to identify distinct

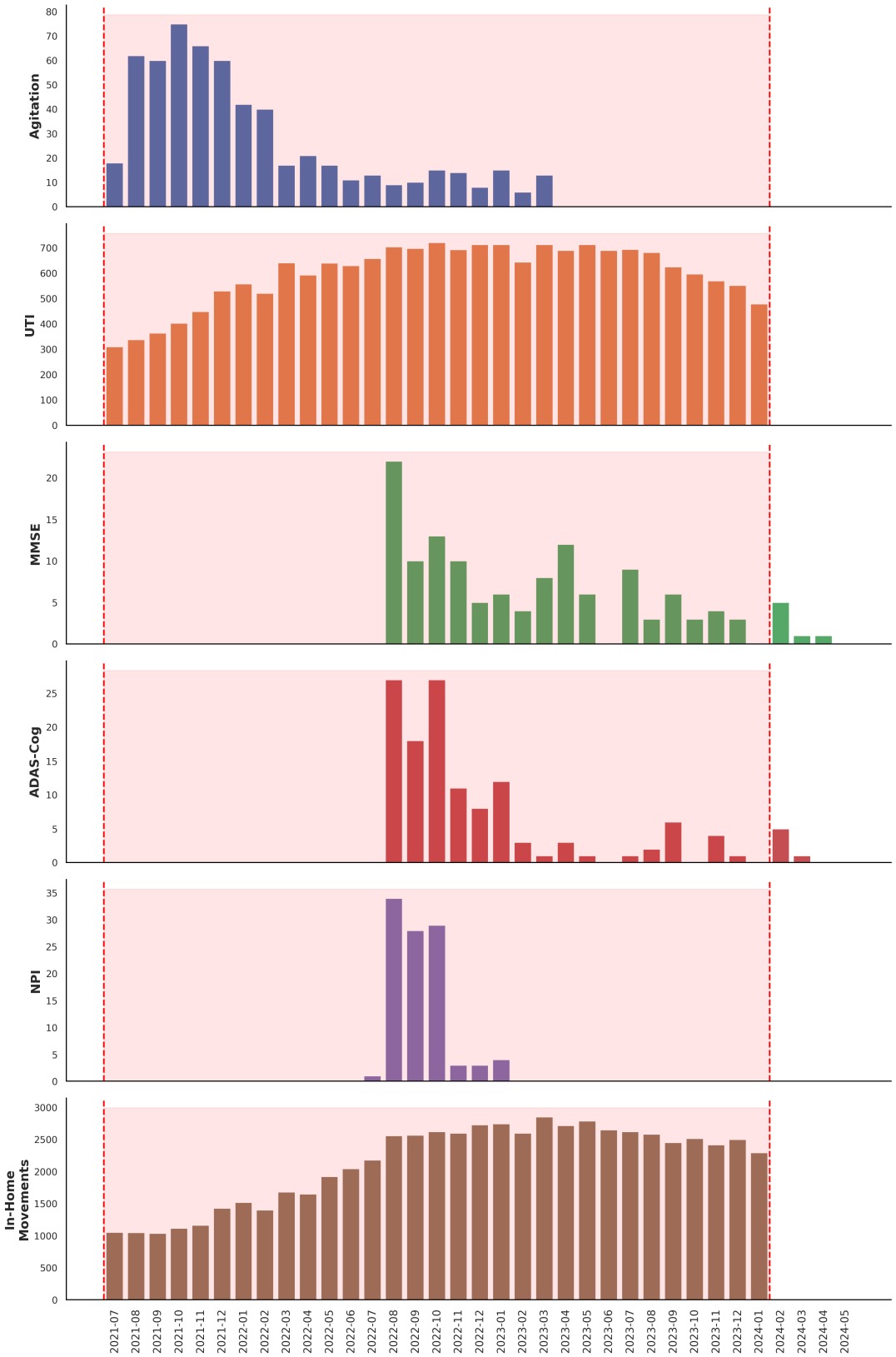


Figure 8: Monthly Distribution of Records

latent states. Finally, Transition Pattern Analysis defines a transition matrix across latent states and applies the PageRank algorithm to evaluate state importance, yielding

a compressed representation of patient behavior patterns that enhances interpretability and supports personalized intervention strategies.



Figure 9: Encoding model pipeline

4.2 Detailed Synthetic Dataset Generation Encoding Process

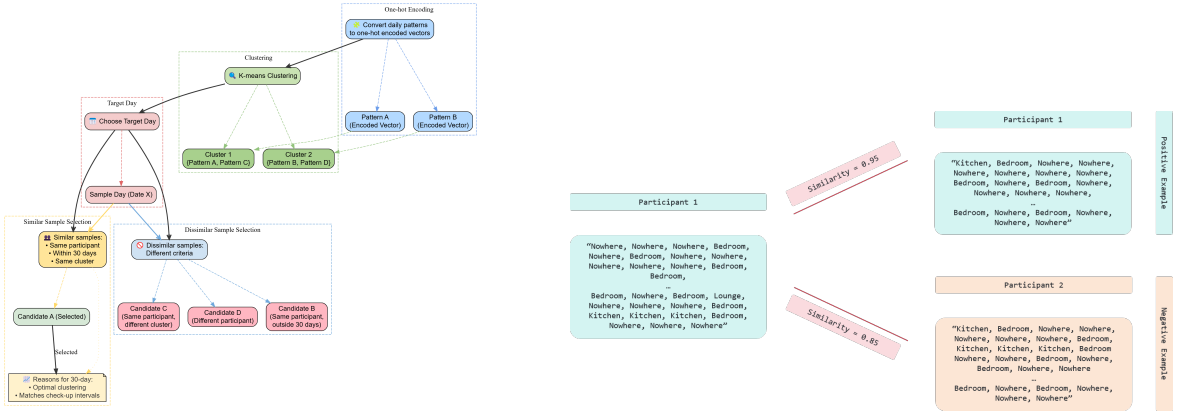
The data set used for home activity monitoring consists of synthetically generated participant movement records that fully replicate the structure and scale of the original empirical data set. This synthetic data set was generated using the encoded representations obtained from the previous stage of model training, using a retrieval-augmented generation (RAG) approach to ensure coherence with real behavioral patterns. The generated movement sequences follow the same format as the original dataset while maintaining entirely virtual patient mobility trajectories.

The original data set included motion data collected from passive infrared sensors installed at various locations in the homes of individuals with dementia, along with sleep monitoring pads placed under their mattresses. These sensors detected motion within predefined ranges and angles, allowing for a detailed record of participant movements across various areas, including lounges, kitchens, hallways, bedrooms, and bathrooms. Additionally, diagnostic data, including dementia diagnoses, cognitive assessments (such as MMSE and ADAS-Cog scores), and demographic details, were available for 134 participants.

For synthetic data generation, the test set was used to extract coding vectors from the fine-tuned language model. The movement trajectories for the synthetic cohort were generated based on prior seven-day movement sequences, conditioned on patient-specific health indicators (such as UTI status, cognitive scores, anxiety, and depression levels). The RAG-based approach retrieved the most similar historical trajectories from the embedding vector database and incorporated these references into the generative model, ensuring that the generated sequences accurately reflected the behavioral trends observed in real patients.

The final synthetic dataset maintains the same structure and scale as the empirical dataset, with each data point including a participant ID, timestamp (to the second), detected activity location, or sleep pad data indicating bed entry or exit events. This data set, which contains 66,096 participant-day records, ensures statistical comparability while being entirely independent of real patient data.

4.2.1 Positive and Negative Sample Selection in Pre-training



(a) Sample Selection Strategy (b) Positive and Negative Sample Example

Figure 10: Pretrained BERT model training strategy

The process of selecting positive and negative samples during the pretraining stage is designed to capture temporal and contextual similarities in patient movement trajectories. The selection strategy and sample are illustrated in Figure 10(a) and 10(b).

4.2.2 Sample Selection Workflow

The workflow for selecting samples is summarized in the Figure 10(a). The process begins with the selection of a target day (*Date X*), which serves as the anchor for the subsequent sample selection. The selection strategy includes two components:

- **Similar Sample Selection:**
  - Similar samples are identified based on three criteria:
    - *Same participant:* The samples belong to the same individual.
    - *Temporal proximity:* Samples are selected from a time window within 30 days of the target day.

    - *Cluster membership:* The selected samples belong to the same cluster defined by a clustering algorithm K-means applied to the encoded movement patterns.
  - The selected candidates are used as positive examples in the training process.
  - A 30-day window of time aligns with clinical check-up intervals, ensuring meaningful comparisons between patterns.

- **Dissimilar Sample Selection:**
  - Dissimilar samples are chosen to maximize the contrast between the target day and the comparison samples.
  - These samples include:
    - *Same participant, different cluster:* Movement patterns of the same individual but belonging to different behavioral clusters.
    - *Different participant:* Samples from other individuals that exhibit different movement patterns.
    - *Outside 30-day window:* Samples temporally distant from the target day.
  - Dissimilar samples serve as negative examples during training.

4.2.3 Positive & Negative Window Selection Process

Given the unlabeled nature of our temporal data, we adopt a group-based contrastive sample selection approach for model training. This method exploits the inherent structure of the data to construct meaningful positive and negative sample pairs.

We conducted a hyperparameter optimization analysis to validate our initial assumptions and determine the most suitable parameters to maximize the separation of vector embeddings. By analyzing K-means clustering results in various configurations, our objective was to identify the optimal setup that enhances the distinction among vector representations while mitigating potential biases.

The 30-day window for positive sample selection was chosen for two reasons: (1) The clustering of K-means of encoded vectors achieved optimal results with this interval, as shown in Table 2; (2) many patients undergo regular monthly health checks, such as urine tests, making this time frame clinically relevant.

Table 2: Silhouette scores under different models and parameter settings

Model	Parameters	Silhouette scores			
		4	5	6	7
MiniLM-L12-v2	7days	0.459	0.451	0.431	0.413
	30days	**0.554**	**0.554**	**0.554**	**0.542**
	180days	0.437	0.429	0.370	0.407
BAAI/bge-small-v1.5[43]	30days	0.459	0.425	0.473	0.473
	no tune	0.173	0.165	0.181	0.170

4.2.4 Embedding and Clustering

To facilitate sample selection, movement trajectories are converted into one hot encoded vectors and processed using a K-means clustering algorithm. Each cluster represents a distinct movement pattern (e.g., *Pattern A*, *Pattern B*), enabling efficient identification of similar and dissimilar samples.

It is crucial to acknowledge a potential source of data leakage introduced at this stage. The behavioral clustering, which underpins the contrastive learning sample selection, is calculated based on movement patterns that may inherently correlate with clinical outcomes (such as UTIs, ADAS-Cog scores, or agitation levels). Consequently, the resulting embeddings are constructed with implicit clinical information. This means the representations are not strictly 'blind' to the clinical labels they are later used to predict, which may partially explain their high efficacy in downstream classification tasks. While this creates a potential bias, it also reflects the real-world entanglement of behavioral phenotypes and clinical states in dementia.

This positive and negative sample selection strategy ensures that the pre-training process leverages meaningful contrasts between movement trajectories, enhancing the model's ability to capture temporal and behavioral nuances in patient data.

4.2.5 Pre-Training Process

Pre-training embeddings are generated using our base model through self-supervised training with the sampling strategy described above. For each training cycle, a random subset of 50,000 triplets was drawn from the full dataset, processed in groups of 256 instances at a time. To assess sentence-level representations, we employed a triplet-based objective function, relying on the computation of Manhattan distance to quantify the dissimilarity between the encoded vectors of anchor items and their paired positive and negative counterparts. Manhattan distance was chosen due to its effectiveness in working with sparse data, its relatively low computational overhead, and its relevance in handling discrete feature spaces. The optimization process used AdamW [44], with a step size of 2×10^{-5} and a regularization term (weight decay) set to 0.01. In addition, a learning rate scheduler with a linear warm-up phase was applied, gradually increasing the rate over the initial 10,000 steps. The training of the large-language model took place on an NVIDIA A100 GPU, with each complete pass through the dataset requiring roughly 444.19 seconds.

4.3 PageRank Iteration for a Single Patient

4.3.1 Model Definition

We compute the PageRank scores and entropy values for individual patients based on their embeddings and cluster labels. This involves defining a transition matrix using distances between clusters and then applying the PageRank algorithm.

4.3.2 Transition Matrix Construction

Given:

- X: Patient embeddings $(n\_samples, 2)$.

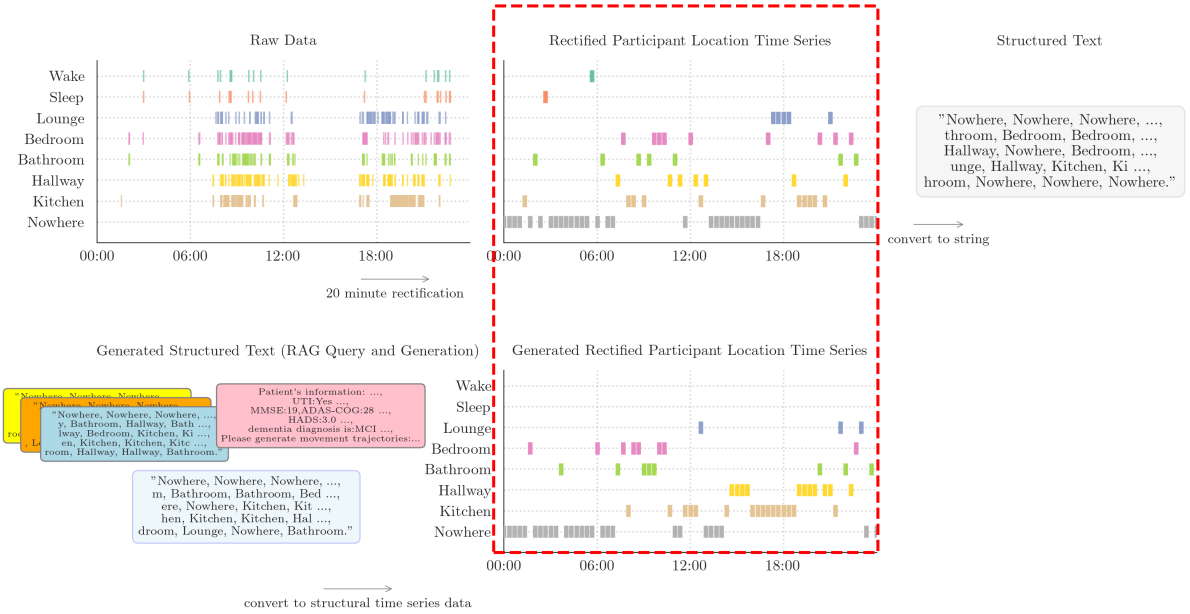


Figure 11: Data Processing and Synthetic Data Generation Visualization

- y: Cluster labels for each sample $(n\_samples,)$.
- $num\_clusters$: Total number of clusters.
- $threshold$: Distance threshold for defining transitions.

The transition matrix $\mathbf{T}$ is computed as:

$$T_{ij} = \frac{\sum_{k \in C_i} \sum_{l \in C_j} \mathbb{1}_{\{d(k,l) \leq \text{threshold}\}}}{\sum_{l \in C_i} \sum_{m \in C_j} \mathbb{1}_{\{d(l,m) \leq \text{threshold}\}}} \tag{1}$$

where:

- C_i and C_j represent samples in clusters i and j, respectively.
- $d(k, l)$ denotes the distance between samples k and l.
- $\mathbb{1}_{\{\cdot\}}$ is an indicator function that returns 1 if the condition is met, otherwise 0.

4.3.3 PageRank Computation

The PageRank vector $\mathbf{p}$ is computed iteratively as:

$$\mathbf{p}^{(t+1)} = \frac{1 - \alpha}{num\_clusters} + \alpha \mathbf{T}^\top \mathbf{p}^{(t)} \tag{2}$$

where α is the damping factor (typically 0.85), and $\mathbf{T}^\top$ is the transpose of the transition matrix. The process continues until convergence:

$$\|\mathbf{p}^{(t+1)} - \mathbf{p}^{(t)}\|_1 < \text{tol} \tag{3}$$

where tol is the predefined tolerance threshold.

4.3.4 Algorithm Summary

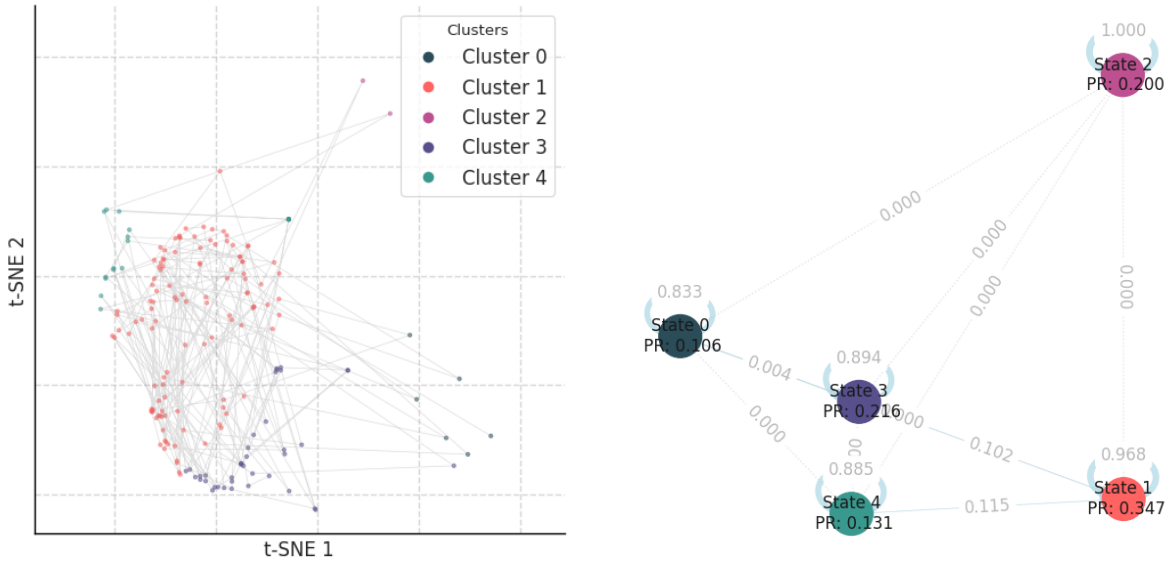
Algorithm 1 outlines the implementation steps of the PageRank model.

Algorithm 1 PageRank for a Single Patient

Input: X, y, $num\_clusters$, $threshold$, α, max_iter, tol
Output: $\mathbf{T}$, $\mathbf{p}$
Initialize transition matrix $\mathbf{T}$ with zeros
for each cluster i **do**
  for each cluster j **do**
    Compute distances between samples in clusters i and j
    Update $\mathbf{T}_{ij}$ based on the distance threshold
  end for
end for
Normalize transition matrix $\mathbf{T}$
Initialize PageRank vector $\mathbf{p}$ uniformly
for iteration $t = 1$ to max_iter **do**
  Compute new PageRank vector $\mathbf{p}^{(t+1)}$
  if convergence condition met **then**
    Break
  end if
end for
Compute final PageRank matrix $\mathbf{P}_{\text{rank}}$

4.3.5 Algorithm Visualization

Figure 12 illustrates the PageRank state generation process.



(a) Latent State Trajectory

(b) Stationary Distribution

Figure 12: Visualization of the PageRank computation process. Panel (a) illustrates the patient's movement trajectory projected into a 2D t-SNE latent space, where clusters represent distinct behavioral patterns. Panel (b) visualizes the computed PageRank values, where the size of each node corresponds to the stationary probability (importance) of that state in the patient's daily routine.

5 Methodology for Building A Latent Representation Model for in-Home monitoring Data

This research proposes a foundation model for medical multimodal data, aiming to represent the complex temporal behavior of PLWD through an interpretable latent structure. It combines self-supervised depth encoding with graph theory-based state modeling and generative data augmentation to provide a tool for personalized clinical prediction, digital twin generation, and the evaluation of its effects. This unified approach addresses model interpretability and data privacy issues, supporting large-scale, low-risk AI deployment in healthcare.

5.1 Preprocessing for Raw Data from Sensors

Our framework began with pre-processing raw temporal data to remove noise and ensure consistency. For encoding, we employed the all-MiniLM-L12-v2 model [45]. We selected this specific architecture because it provides an optimal trade-off between inference latency and semantic embedding quality. With approximately 33 million parameters, it achieves competitive performance on the Massive Text Embedding Benchmark (MTEB) while remaining computationally efficient. This efficiency is critical for our study, as processing the high-frequency dataset of over 24 million sensor records requires a model that minimizes computational overhead without compromising the capture of intrinsic temporal dependencies. To optimize its performance for our specific dataset, we fine-tuned the model using its pre-trained weights. The processed temporal data were then transformed into 384-dimensional vector representations, effectively preserving intrinsic temporal dependencies and patterns.

5.2 Data Extraction and Engineering

Given the unlabeled nature of our temporal data, we utilized a cluster-based contrastive sample selection method with triplet loss for training and evaluation, as illustrated in Figure 10(a). Further details on the language model training process are provided in the Appendix.

To interpret the resulting high-dimensional embeddings, we used t-SNE [46] to project the data into a more manageable 2D space. Subsequently, K-means clustering was applied to identify distinct latent states within these projections. Since data points are temporally ordered, this 2D projection allowed for a clear visualization of the movement patterns of each participant within the latent activity space.

Finally, we constructed a transition matrix between the identified latent states and applied the PageRank algorithm [27] to rigorously analyze transition patterns. Unlike standard Markov transition matrices that focus solely on immediate step probabilities, PageRank evaluates the global importance of nodes within the graph structure, converging to a stationary distribution. In the context of dementia monitoring, this stationary distribution mathematically represents a patient's 'routine' or most stable behavioral states. By quantifying the weight of these states, we can rigorously detect deviations from established routines, which is a key clinical indicator of cognitive decline or acute events. Further details on this algorithm are available in the Appendix. This study presents a scalable approach for analyzing large amounts of temporal data related to patient movement behavior. For context, a single participant with complete continuous

data collected over a 3-month period, with passive infrared (PIR) sensors that record data in seconds, would generate approximately $3 \times 30 \times 86,400 = 7,776,000$ sparse records. Such a data volume is infeasible for direct human-expert analysis and would demand extensive computational resources from most traditional machine learning models. Our method directly addresses this challenge by transforming this high-dimensional dataset into a low-dimensional, semantically interpretable representation. Specifically, through our two-stage encoding and PageRank-based dimensionality reduction, we compressed the data into a latent state vector of length five. Each element in this vector is not only computationally efficient, but also carries interpretable semantic information, facilitating transparent and personalized insights into patient behavior dynamics. This approach exemplifies the application of low-rank structures to manage and interpret complex temporal data sets within the field of AI in healthcare care, supporting personalized intervention strategies, as shown in Figure 13.

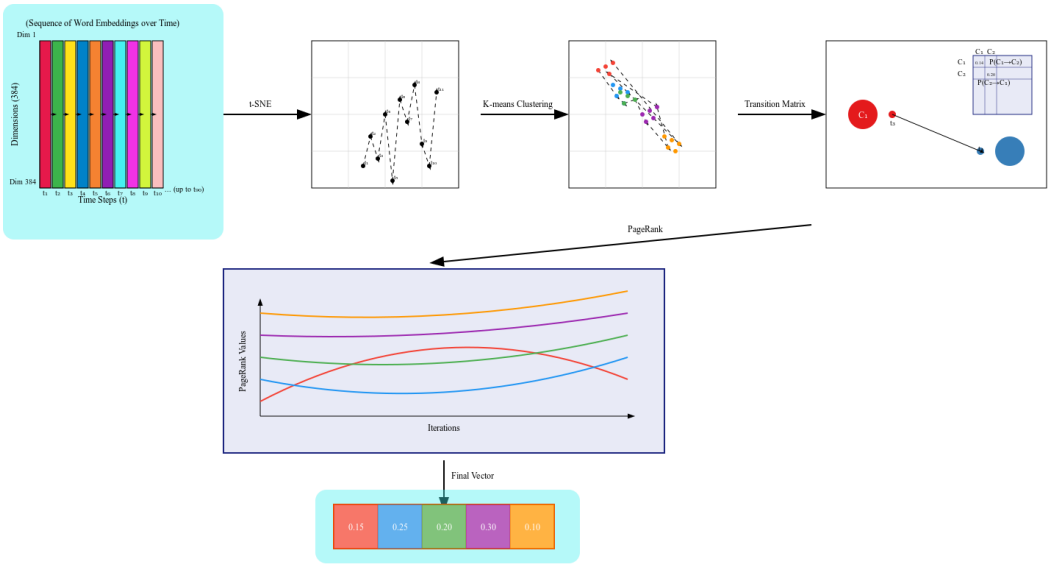


Figure 13: Encoding Model Pipeline

5.3 Data Decoding

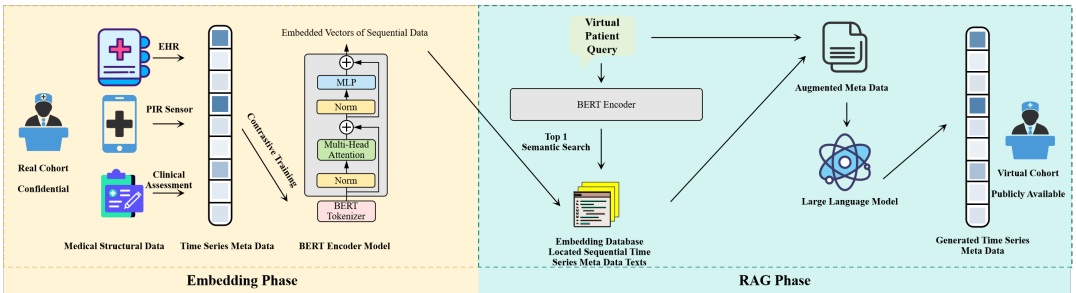


Figure 14: Decoding Model Pipeline

The proposed pipeline, as shown in Figure 14, integrates sequential data embedding, retrieval-augmented generation (RAG), and large language models (LLMs) to construct

privacy-preserving virtual cohorts while maintaining temporal fidelity and clinical relevance.

5.3.1 Retrieval Augmented Generation

The RAG Phase employs a Retrieval-Augmented Generation framework to generate augmented meta-data and synthetic cohorts. The key steps are as follows:

- **Virtual Patient Query:** A user-defined query specifying the characteristics of the synthetic cohort to be generated.
- **Semantic Search:** The BERT encoder is used to generate embeddings for the virtual query, which are then matched against a vector database to retrieve the most similar meta-data texts from time series.
- **Large Language Model (LLM):** The data extracted is combined with the query and processed through an LLM to generate augmented meta-data. This ensures clinical plausibility by incorporating prior medical knowledge.
- **Virtual Cohort Generation:** The augmented meta-data is transformed into a structured virtual cohort, which maintains the statistical properties of the original data while aiming to provide a layer of data anonymization. However, since this process relies on retrieving actual patient trajectories, the generated cohorts may exhibit high fidelity to real individuals, necessitating further scrutiny regarding privacy preservation.

5.4 Downstream Applications and Validation

The virtual cohorts generated are designed to support downstream tasks such as modeling disease trajectory, clinical decision support, and personalized treatment planning. The pipeline ensures data privacy while enabling the broader use of clinically relevant data sets for research and development.

This pipeline is built to address the dual challenge of privacy preservation and the need for high-fidelity synthetic data in healthcare applications.

This workflow demonstrates the integration of real-world data rectification, structured text generation, and RAG-based synthetic data generation. The resulting synthetic data maintain the statistical and behavioral properties of the original dataset while ensuring privacy preservation.

To evaluate the quality and fidelity of the generated data, a comprehensive set of unsupervised evaluation metrics was used, as illustrated in Figure 15. These methods analyze the embedding vectors derived from movement data (PIR sensor recordings) and electronic health records (EHRs) using a pretrained language model. Evaluation metrics include cosine similarity, principal component analysis (PCA) visualization, Jensen-Shannon (JS) distance, and Fréchet inception distance (FID) score.

Cosine Similarity quantifies the alignment between real and generated data embeddings, with higher values indicating stronger similarities in vector representations. PCA Visualization offers a qualitative perspective by projecting high-dimensional embeddings into a two-dimensional space, allowing for visual comparisons of the distribution and clustering patterns. The Jensen-Shannon distance measures the divergence between the probability distributions of real and generated data, with lower distances reflecting a

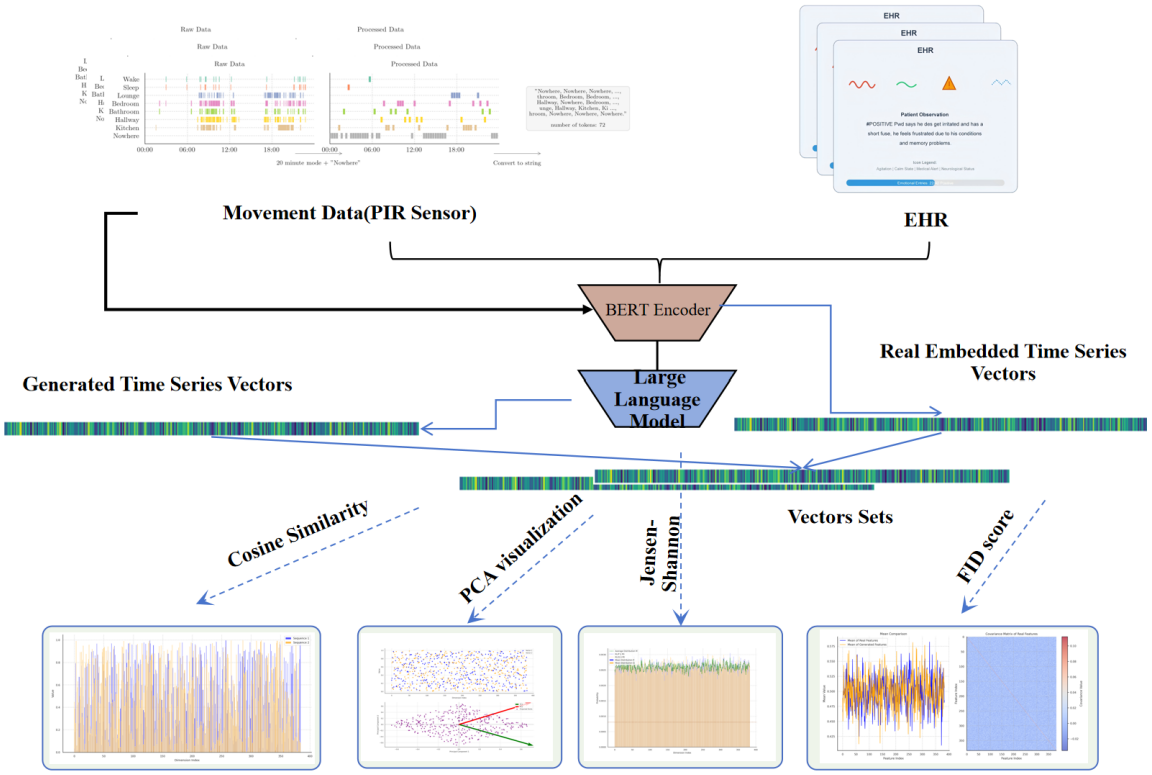


Figure 15: Unsupervised Methods

closer statistical match. Finally, the FID score assesses both global (mean) and local (variance) characteristics of the data distributions, with lower scores signifying better fidelity.

This evaluation framework combines statistical and structural analyses to ensure that the generated data not only replicate the real data's statistical properties but also preserve its contextual and temporal patterns. By integrating these complementary metrics, we establish a robust methodology to assess the performance of generative models in healthcare applications, supporting their utility for downstream tasks and clinical decision-making.

The generated virtual time series data sets were utilized to address a variety of downstream healthcare tasks, as shown in Figure 16. These tasks included predicting urinary tract infections (UTIs), cognitive and neuropsychiatric assessments (ADAS-Cog, NPI, MMSE) and episodes of agitation. Two types of neural networks, LSTMs and Transformer models, were employed to process the sequential data. LSTMs captured long-range temporal dependencies for diagnostic predictions, while Transformers leveraged multihead attention mechanisms to extract complex temporal and contextual features. The results were categorized into classification tasks, such as UTI detection and prediction of agitation, and regression tasks for clinical scale scores. This framework demonstrates the utility of synthetic datasets in supporting diagnostic, prognostic, and behavioral predictions, enabling scalable, privacy-preserving research in diverse clinical applications.

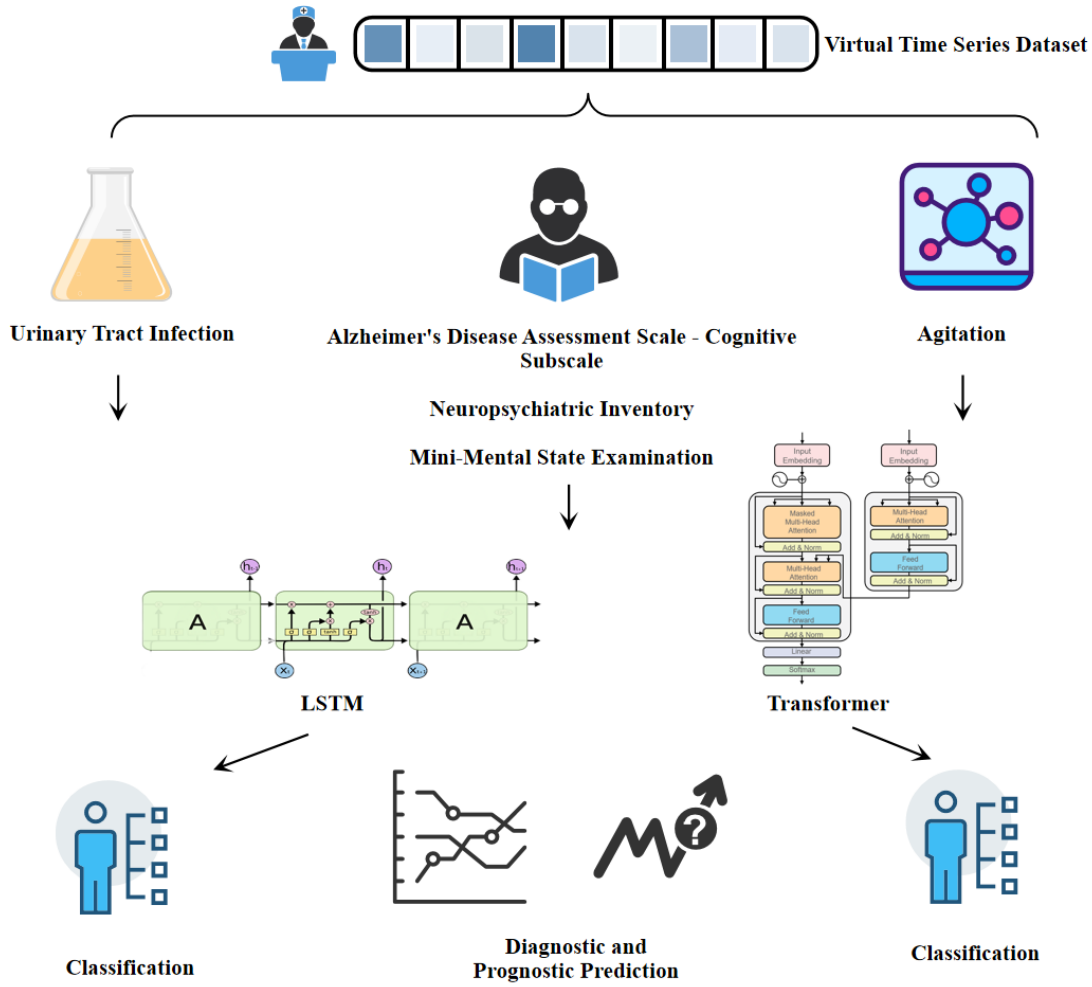


Figure 16: Downstream Tasks Pipeline

6 Exploratory Analysis and Downstream Model Development

This chapter provide more detailed insights into the foundation model development and evaluation.

6.1 Encoder Model

Applying K-means clustering to the text embeddings of the test set, we identified an optimal solution with five clusters, indicating five latent states underlying single-day, single-participant behavioral patterns. Figure 17 visualizes these clusters after the reduction in dimensionality of t-SNE. By analyzing the transformation of individual vectors in this 2D space, we can trace participants' behavioral trajectories within the embedding space (see the appendix for additional visualizations). In collaboration with clinical experts, we further examine the semantic meaning of these clusters and their associations with patient medical characteristics.

More importantly, by integrating the random walk model and PageRank algorithm with clinical expertise and diagnostic results, we can quantitatively evaluate the deeper

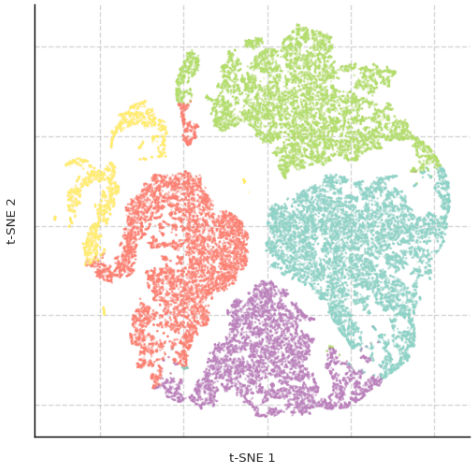


Figure 17: T-SNE for embedded daily movement strings in the test set

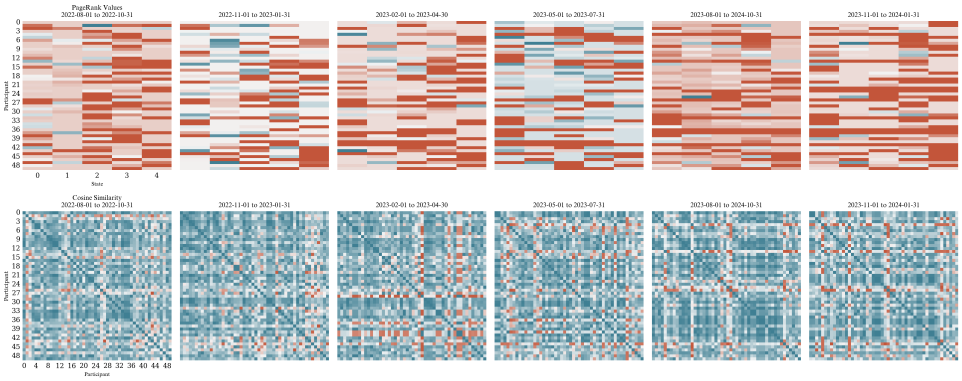


Figure 18: Multi-period participant deep state vector and similarity

semantic meanings embedded within vector clusters or latent states. Figure 18 presents a multiperiod heatmap analysis, segmented in five time intervals to visualize patient behavior dynamics. The top row of heatmaps illustrates the PageRank values in different latent states (1-5) for each participant over time, with each interval representing a period of three months. The intensity of color in each cell reflects the PageRank value for a particular state and participant, highlighting shifts in dominant latent states and revealing periodic patterns in behavior. This aggregation reduces the high-dimensional behavioral data to a low-rank, interpretable representation. The bottom row of the heatmaps shows the cosine similarity matrices between the participants for each respective time interval. These matrices indicate the degree of similarity between participants' behaviors within each period, with higher similarity values represented by darker shades. The block patterns visible in some intervals suggest clustering tendencies among participants with similar behavioral patterns. Heatmap analysis reveals distinct behavioral patterns among participants, with certain individuals demonstrating notable stability and periodicity in their daily activity states. For example, several participants consistently show high PageRank values for specific latent states (such as state 2 or 3) across multiple time intervals. This stability may indicate regular lifestyle patterns or consistent routines, which can be critical in predicting daily behavior and planning personalized interventions. In contrast, some participants exhibit significant shifts in their PageRank values between periods, particularly around seasonal transitions or specific time frames (e.g., August-October 2022 and

May-July 2023). These variations may reflect seasonal effects or environmental changes that influence patient behavior, underscoring the need to account for time-related factors when designing personalized monitoring and intervention strategies. The cosine similarity heatmaps reveal clusters of participants who show high similarity in behavior during certain intervals (for example, May-2023 and August 2023-January 2024), forming identifiable groups. This similarity may suggest shared behavioral characteristics or similar health conditions between these participants, possibly due to common lifestyle factors, environmental influences, or comparable stages in disease progression. These group-level patterns provide information on cohort-based behavioral dynamics, supporting the development of individualized and group-based healthcare management strategies.

Based on clustering and SHAP analyses of PageRank-derived states and cognitive metrics (see the appendix), distinct behavioral and cognitive profiles emerge across clusters. Each PageRank state, analyzed through correlation and SHAP values, reveals unique associations with cognitive scores (MMSE, ADAS-Cog), age, and mood-related metrics (HADS Depression and Anxiety). By reducing the complex vector matrix to a simplified (1,5) vector (see Appendix), we can further explore the semantic characteristics of each state.

Looking ahead, once a unique deep vector is generated for each participant, these states can be used to predict MMSE and ADAS-Cog scores, along with their rate of change. Detailed methodologies and implementation specifics are provided in the Appendix for further reference.

6.2 Cognitive Status Prediction: Component Ablation Study

To validate that the deep states from our model encapsulate semantic information regarding cognitive status, we conducted a component ablation study. This study systematically removes different parts of the feature input pipeline to evaluate the specific contribution of our proposed representation learning module (the PageRank-derived 'State' vectors) against traditional methods.

The baseline model, a Random Forest, was trained using conventional features: partial clinical metrics and simply aggregated movement trajectories. By comparing the performance of the model with and without the 'State' vectors, we can rigorously justify the inclusion of the encoder and PageRank steps in our architecture.

To evaluate the impact of generated state features on predictive performance, we consider several feature combinations as ablation targets:

- **Baseline:** Original location count-based features (mean and variance).
- **State Features (Proposed Component):** Representation of patient states (`state1` to `state5`) derived from the proposed encoding pipeline.
- **Characteristics:** Patient-specific characteristics such as age, sex, and diagnosis of dementia.
- **Combined Features:** Various combinations of the above, including all features.

The results in Table 20 demonstrate that isolating the 'State' component (or combining it with characteristics) consistently yields lower MAE errors compared to the Baseline features alone. This confirms that the proposed representation learning pipeline captures behavioral dynamics that are lost in simple statistical aggregations, justifying the complexity of the two-stage encoding model.

Table 3: Performance metrics for predicting current ADAS-Cog and MMSE with 95% confidence intervals (CI). The values in parentheses represent the lower and upper bounds of the 95% CI.

Model	Metric Feature Set	$\mathbf{MAE_{ADASCOG}}$	$\mathbf{MAE_{MMSE}}$	$\mathbf{RMSE_{ADASCOG}}$	$\mathbf{RMSE_{MMSE}}$
RandomForest	Baseline	11.41 (9.20, 13.65)	4.64 (3.83, 5.50)	11.37 (9.25, 13.64)	4.68 (3.90, 5.50)
	Baseline + Characteristics	10.81 (8.63, 13.10)	4.55 (3.76, 5.43)	10.85 (8.76, 13.24)	4.59 (3.73, 5.51)
	Characteristics	10.62 (8.27, 13.08)	4.34 (3.39, 5.36)	10.65 (8.24, 13.13)	4.33 (3.34, 5.30)
	State	10.46 (8.31, 12.63)	4.07 (3.09, 5.03)	10.41 (8.25, 12.54)	4.09 (3.20, 5.09)
	State + Baseline + Characteristics	10.50 (8.43, 12.69)	4.37 (3.58, 5.25)	10.62 (8.39, 12.84)	4.39 (3.50, 5.30)
	State + Characteristics	10.31 (8.21, 12.79)	4.19 (3.32, 5.06)	10.34 (8.33, 12.55)	4.19 (3.30, 5.13)
Ridge	Baseline	12.77 (10.23, 15.39)	5.15 (4.11, 6.31)	12.71 (10.24, 15.22)	5.10 (4.00, 6.30)
	Baseline + Characteristics	12.04 (9.67, 14.43)	5.03 (4.01, 6.17)	11.93 (9.66, 14.34)	5.05 (4.08, 6.16)
	Characteristics	10.34 (8.33, 12.47)	4.35 (3.51, 5.34)	10.31 (8.25, 12.53)	4.34 (3.47, 5.37)
	State	**9.73** (7.31, 12.20)	**3.81** (2.81, 4.89)	**9.83** (7.38, 12.42)	**3.81** (2.74, 4.93)
	State + Baseline + Characteristics	13.20 (10.53, 16.14)	5.49 (4.40, 6.53)	13.23 (10.73, 16.01)	5.49 (4.39, 6.59)
	State + Characteristics	10.18 (7.84, 12.87)	4.27 (3.26, 5.32)	10.15 (7.69, 12.70)	4.25 (3.18, 5.42)

6.3 Decoder Model

Please note that the mathematical calculations here are performed on vectors obtained by directly encoding the textual representations of real in-home movement data and the generated virtual cohort in-home movement data using these base models.

6.3.1 Unsupervised Methods Evaluation

Model	Cosine Similarity	Average Jensen-Shannon Distance	FID Score	Precision/Recall(k=1)	
				Precision	Recall
all-MiniLM-L12-v2	0.9045	0.3264	0.0212	0.9251	0.9248
all-MiniLM-L6-v2	0.5032	0.3967	0.9224	0.9324	0.9280
paraphrase-MiniLM-L12-v2	0.5129	0.3447	27.8465	0.9335	0.9323
stsb-roberta-base-v2	0.8206	0.2784	9.8821	0.9187	0.9176
all-mpnet-base-v2	0.8504	0.2623	0.0324	0.9190	0.9281
average	**0.7183**	**0.32857**	**7.7349**	**0.93147**	**0.92657**

Table 4: Unsupervised Evaluation Results on Synthetic Dataset

To rigorously justify the selection of the embedding backbone, we conducted a comparative analysis of five distinct transformer-based models, as listed in Table 4. These models were chosen to span a diverse range of parameter sizes (from 22M in MiniLM to 110M in MPNet) and architectures (including distilled BERT, RoBERTa, and MPNet). This selection allows us to investigate the trade-off between computational complexity and the fidelity of the generated synthetic data (measured by FID and JS Distance), ensuring that the final foundational model is not only accurate but also scalable for real-world healthcare deployment.

Table 4 summarizes the performance of five embedding models in multiple evaluation metrics, including cosine similarity, average Jensen-Shannon (JS) distance, Fréchet Inception Distance (FID) and precision and recall of k-Nearest Neighbors (k-NN) at $k = 1$.

The `all-MiniLM-L12-v2` model achieved the highest cosine similarity (0.9045) and the lowest FID score (0.0212), indicating its superior ability to preserve characteristic similarity and distributional fidelity. Furthermore, the `paraphrased MiniLM-L12-v2` model excelled in k-NN precision and recall, achieving values of 0.9335 and 0.9323, respectively, demonstrating its robustness in aligning generated and real data in the feature space.

For the average JS distance, the `all-mpnet-base-v2` model outperformed others with the lowest value (0.2623), suggesting a better generalization across distributions. In contrast, the `all-MiniLM-L6-v2` model displayed suboptimal performance in cosine similarity (0.5032) but performed relatively well in k-NN precision (0.9324) and recall (0.9280).

On average, the embedding models achieved a cosine similarity of **0.7183**, an average JS distance of **0.32857**, an FID score of **7.7349**, and k-NN precision and recall values of **0.93147** and **0.92657**, respectively. These results demonstrate that embedding models can effectively balance feature preservation, distributional alignment, and generalization, with each model showing strengths in specific metrics. Figure 19 presents the results of the PCA comparison of the PLWD movement trajectory strings, encoded using `all-MiniLM-L12-v2`, for both empirical (real) and virtual (generated) cohorts.

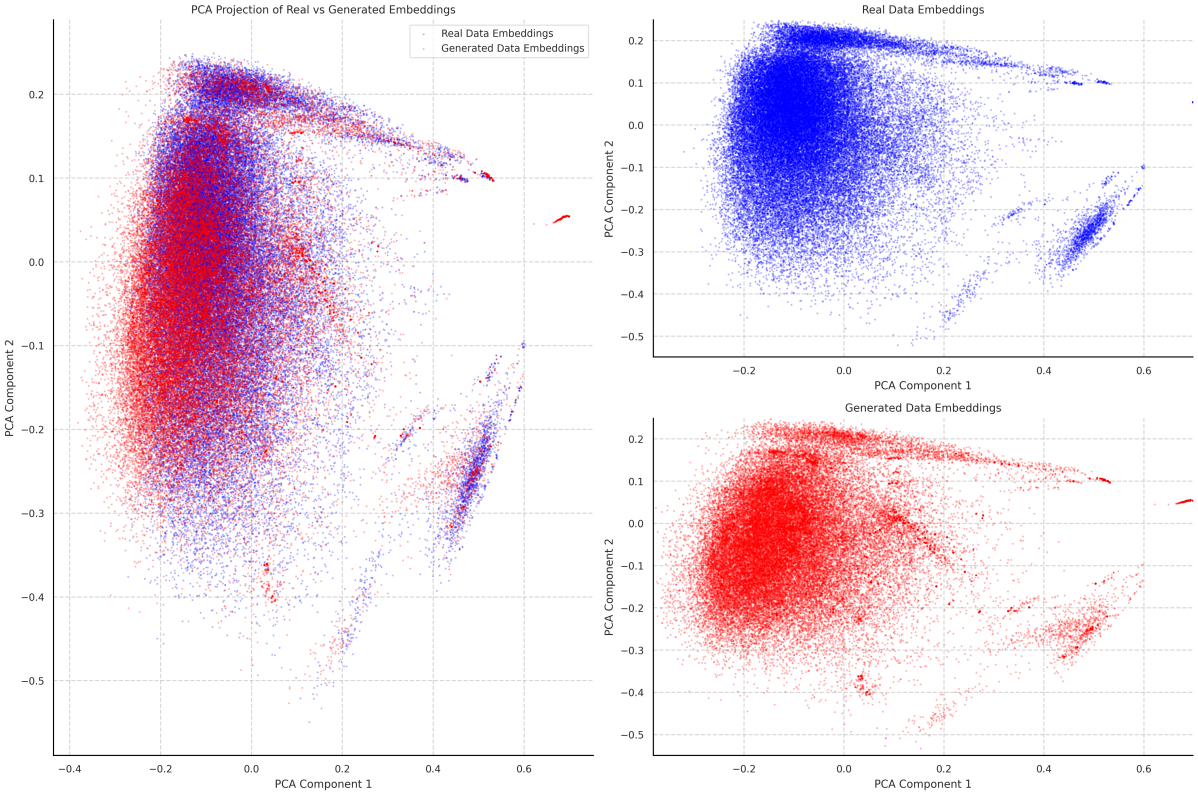


Figure 19: PCA Visualization of Real vs. Generated PLWD Movement Trajectory Embeddings (all-MiniLM-L12-v2)

Comparative analysis of five-single embedding models reveals distinct performance patterns across multiple evaluation metrics, as shown in Figure 20. Each metric is normalized to a range between 0 and 1, with values close to 1 representing better performance. For metrics where lower values are desirable (e.g. Jensen-Shannon divergence and FID score), the normalization was adjusted such that smaller values are mapped to higher scores (i.e., inverted normalization). This ensures consistency in interpretation, where a higher normalized score consistently reflects better model performance. The larger models, specifically `stsb-roberta-base` and `all-mpnet-base`, demonstrate superior parameter efficiency with 125M and 110M parameters, respectively, significantly exceeding their smaller counterparts. These models also maintain consistently higher

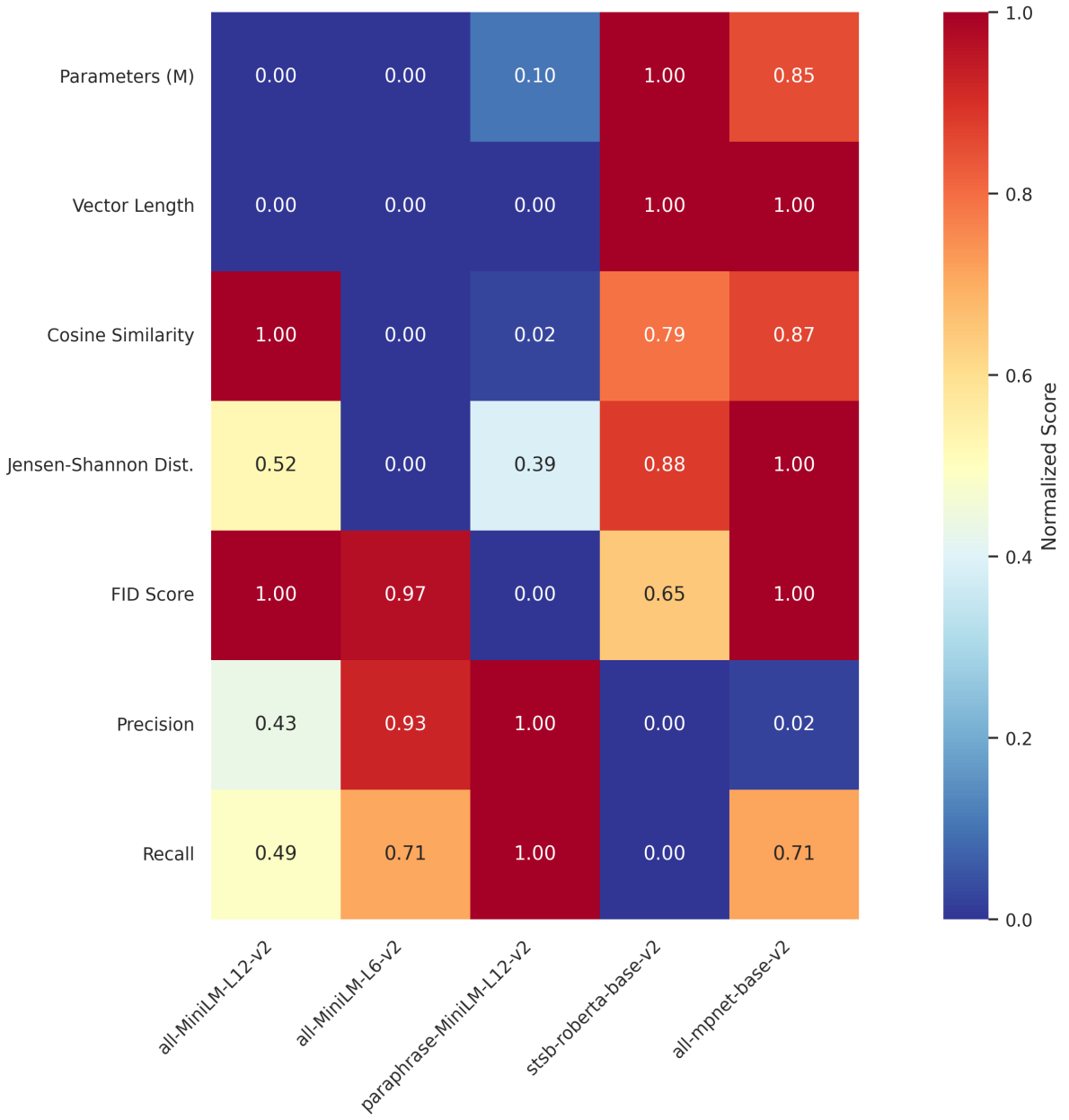


Figure 20: Model Comparison Heatmap

vector dimensionality (768D), which may contribute to their enhanced representational capacity. Notably, `all-MiniLM-L12-v2` achieves remarkable efficiency, displaying competitive performance in semantic similarity tasks (cosine similarity: 0.9045) despite its relatively compact architecture (22.8M parameters).

In terms of semantic preservation, as measured by the Jensen-Shannon distance and FID score, `all-mpnet-base` exhibits the most stable performance (JS distance: 0.2623, FID: 0.0324), indicating superior preservation of the original semantic space. The precision and recall metrics show minimal variance across all models (range: 0.91-0.93), indicating robust performance in downstream classification tasks. Interestingly, the `paraphrase-MiniLM-L12-v2` variant, despite its moderate parameter count (33.5M), achieves the highest precision (0.9335), demonstrating that architectural design can some-

times outweigh the size of the model in specific evaluation contexts.

These findings suggest that, while larger models generally offer better performance across most metrics, careful architectural choices in smaller models can yield competitive results for specific applications. This observation has significant implications for deployment scenarios where computational resources are limited, as smaller models maintain reasonable performance while requiring substantially fewer resources.

6.3.2 Downstream Task Performance

For the division of the training and test sets, to avoid data leakage, we assign 80% of individual patients and all their corresponding records to the training set, while the remaining 20% of individual patients and their records are assigned to the test set. For details on the time series data used as model input, please refer to the relevant content in the Appendix A.11.

Table 5: Model Performance for Different Down Stream Tasks

Embedding Model	Prediction Model	UTI[1]		ADAS-Cog[2]		NPI[2]		MMSE[2]	
		Empirical (95% CI)	Virtual (95% CI)	Empirical (95% CI)	Virtual (95% CI)	Empirical (95% CI)	Virtual (95% CI)	Empirical (95% CI)	Virtual (95% CI)
all-MiniLM-L12-v2	LSTM	0.756 (0.718, 0.794)	0.780 (0.746, 0.814)	12.4120 (11.3538, 13.4702)	12.4130 (11.3515, 13.4744)	9.3089 (7.9277, 10.6902)	9.3089 (7.9277, 10.6902)	4.6482 (3.8383, 5.4581)	4.6620 (3.8308, 5.4931)
	Transformer	0.795 (0.765, 0.818)	0.906 (0.885, 0.925)	12.1474 (11.1188, 13.1760)	12.2773 (11.2130, 13.3415)	9.0521 (7.7405, 10.3636)	8.7584 (7.5048, 10.0120)	4.3071 (3.4983, 5.1158)	4.2711 (3.4580, 5.0843)
all-MiniLM-L6-v2	LSTM	0.732 (0.700, 0.763)	0.727 (0.710, 0.744)	12.4267 (11.3654, 13.4880)	12.4283 (11.3672, 13.4894)	9.3023 (7.9186, 10.6860)	9.3120 (7.9202, 10.7038)	4.6992 (3.8480, 5.5504)	4.6117 (3.7806, 5.4428)
	Transformer	0.758 (0.721, 0.795)	0.714 (0.696, 0.732)	12.2072 (11.1468, 13.2676)	12.0566 (10.9792, 13.1339)	9.1708 (7.8426, 10.4991)	9.0991 (7.7896, 10.4086)	4.3417 (3.5284, 5.1550)	4.2969 (3.4874, 5.1064)
paraphrase-MiniLM-L12-v2	LSTM	0.747 (0.722, 0.772)	0.754 (0.723, 0.784)	12.4304 (11.3654, 13.4954)	12.4331 (11.3750, 13.4913)	9.2591 (7.8908, 10.6273)	9.2300 (7.8759, 10.5841)	4.7325 (3.8760, 5.5891)	4.6442 (3.8041, 5.4843)
	Transformer	0.745 (0.725, 0.765)	0.745 (0.712, 0.777)	12.1681 (11.0871, 13.2491)	12.2563 (11.1901, 13.3225)	9.0078 (7.7180, 10.2976)	8.9560 (7.6937, 10.2182)	4.2984 (3.4771, 5.1198)	4.5245 (3.6465, 5.4025)
stsb-roberta-base-v2	LSTM	0.738 (0.713, 0.764)	0.736 (0.717, 0.756)	12.4358 (11.3786, 13.4931)	12.4385 (11.3746, 13.5024)	9.1877 (7.8481, 10.5274)	9.0835 (7.7871, 10.3799)	4.7241 (3.8937, 5.5545)	4.5519 (3.7438, 5.3600)
	Transformer	0.730 (0.709, 0.750)	0.738 (0.716, 0.761)	12.1552 (11.0985, 13.2119)	12.2071 (11.1463, 13.2678)	9.0107 (7.7374, 10.2840)	9.1334 (7.8680, 10.3987)	4.3504 (3.5382, 5.1627)	4.3180 (3.5104, 5.1256)
all-mpnet-base-v2	LSTM	0.763 (0.743, 0.782)	0.732 (0.709, 0.754)	12.4148 (11.3569, 13.4728)	12.4405 (11.3893, 13.4917)	9.2568 (7.8746, 10.6391)	9.2552 (7.8777, 10.6327)	4.4687 (3.6521, 5.2852)	4.6039 (3.7836, 5.4242)
	Transformer	0.752 (0.725, 0.778)	0.736 (0.720, 0.752)	12.1978 (11.1543, 13.2413)	12.2141 (11.1822, 13.2459)	9.1873 (7.8977, 10.4769)	8.8431 (7.5548, 10.1314)	4.2752 (3.4649, 5.0855)	4.2723 (3.4582, 5.0865)

Table 6: Model Performance for Different Down Stream Tasks (Filtered)

Embedding Model	UTI[1]		ADAS-Cog[2]		NPI[2]		MMSE[2]	
	Empirical (95% CI)	Virtual (95% CI)	Empirical (95% CI)	Virtual (95% CI)	Empirical (95% CI)	Virtual (95% CI)	Empirical (95% CI)	Virtual (95% CI)
all-MiniLM-L12-v2	0.795 (0.765, 0.818)	0.906 (0.885, 0.925)	5.1464	5.7919	7.2881	6.8960	1.4537	1.3832

Please note that in our model input, prior information includes UTI, ADAS-Cog, NPI, and MMSE data for each record, provided that these indicators are available for the specific patient at a given time period. These four indicators are incorporated into the model to generate outputs; therefore, they are evaluated together. Additionally, we have a clinical indicator, agitation, which was deliberately not included as prior information in the model input. This approach allows us to assess the robustness of the model-generated data, particularly for tasks involving unseen variables. Therefore, the classification results for agitation will be presented separately.

Table 5 summarizes the performance of various embedding models and prediction architectures across multiple downstream tasks, including UTI prediction, ADAS-Cog, NPI, and MMSE. For each task, the models were evaluated on both real and synthetic datasets to assess their predictive accuracy, as measured by the mean absolute error (MAE).

Since agitation was not included as prior information in the model input, comparing different base encoding models is not meaningful. According to the results presented in Table 5, the choice of different backbone model combinations does not significantly impact overall model performance. Since the primary focus of this study is not to identify

[1]Metric: Accuracy
[2]Metric: MAE
[1]Metric: Accuracy
[2]Metric: MAE

the optimal backbone combination, we select the most suitable configuration based on the model's capabilities and performance across various metrics. Therefore, we select `all-MiniLM-L12-v2` as the base model and use a Transformer-based architecture as the classification model. The specific parameter settings remain entirely consistent with previous configurations.

Please note that all subsequent ablation studies and performance evaluations of the test set are conducted on the basis of this chosen model combination.

It can be observed that nearly all downstream task performances based on virtually generated cohorts are not inferior to those based on empirical cohorts. In fact, for the selected combinations of base models, many virtual cohorts even outperform their empirical counterparts. For example, in the `all-MiniLM-L12-v2` model with a Transformer prediction model, the synthetic cohort achieved a UTI score of 0.906 (95% CI: 0.885, 0.925) compared to 0.795 (95% CI: 0.765, 0.818) for the empirical cohort, and the ADAS-Cog score was 12.2773 (95% CI: 11.2130, 13.3415) versus 12.1474 (95% CI: 11.1188, 13.1760) in the empirical cohort. These examples indicate that virtual cohorts can serve as reliable—and in some cases even superior—alternatives to empirical data for downstream tasks.

Figure 21, 22, 23 presents the prediagnostic results for three cognition-related assessments, ADAS-COG, MMSE, and NPI, obtained using the optimal model using leave-one-out strategy, focusing on subjects with the highest number of confirmed test results. The figures illustrate the time to testing, displaying the actual test result values alongside the predicted results derived from the real dataset within the previously established framework. In addition, the prediction results for the virtual data set, obtained in the same framework, are presented along with their corresponding confidence intervals. From these three sets of figures, it can be observed that the overall trends between the actual measured values at different follow-up periods (indicated by red or green markers) and the predictions from the two datasets ("Empirical Cohort" and "Virtual Cohort") remain relatively consistent. Notably, the predictions from the "Empirical Cohort" align more closely with the actual values at most time points, suggesting its greater adaptability in capturing individual temporal dynamics. In contrast, the "Virtual Cohort" exhibits a more stable prediction trend, though slight discrepancies are observed in certain patients.

6.4 Classification Risk Analysis

For evaluating the quality of the generated dataset, in addition to the previously mentioned unsupervised mathematical approaches and the average performance across various downstream tasks, we also introduce a more fine-grained method in this section. This approach treats the virtual cohort and the real cohort as two distinct raters for comparison, providing a more detailed assessment of data set quality. In this experiment, I selected the model that performed best on the 10-fold cross-validation training set to evaluate its performance on the test set. Figure 25, 27 presents the pre-diagnostic outcomes for the two classification tasks—UTI and agitation—derived using the optimal model selected from the training set. These analyses were performed on subjects who exhibited the highest number of confirmed test results in the test set. The figures illustrate the time to negativity and positivity for both tests, metrics that, according to our clinical experts, are probably associated with the patient movement trajectory records for the previous seven days. Consequently, the bars in the graphs serve to quantify this effect. Additionally, beneath each predicted negative and positive outcome, the probability of

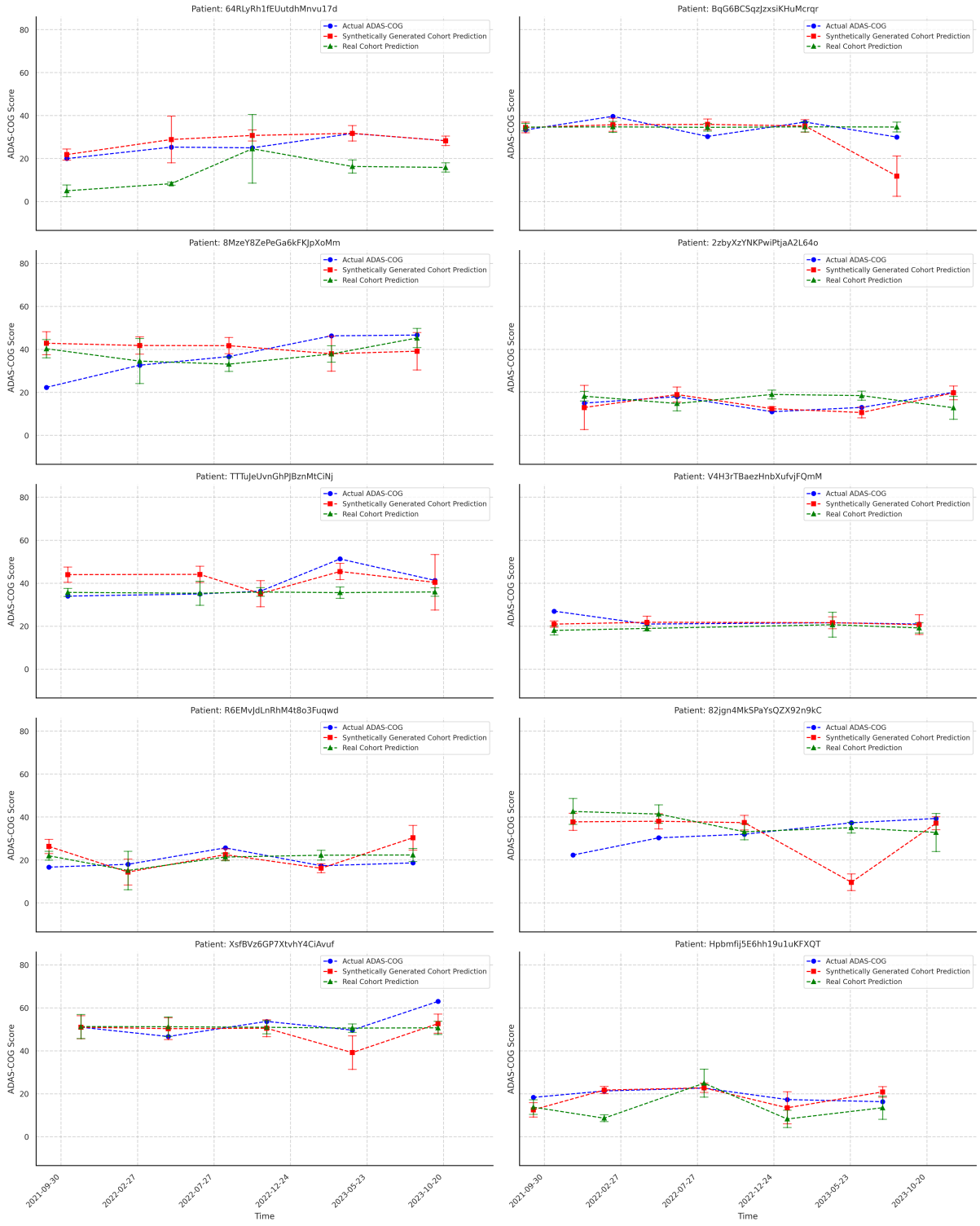


Figure 21: ADAS-Cog Prognostic Results on Test set

the model's verdict is provided, offering a more intuitive and interpretable evaluation of the predictions.

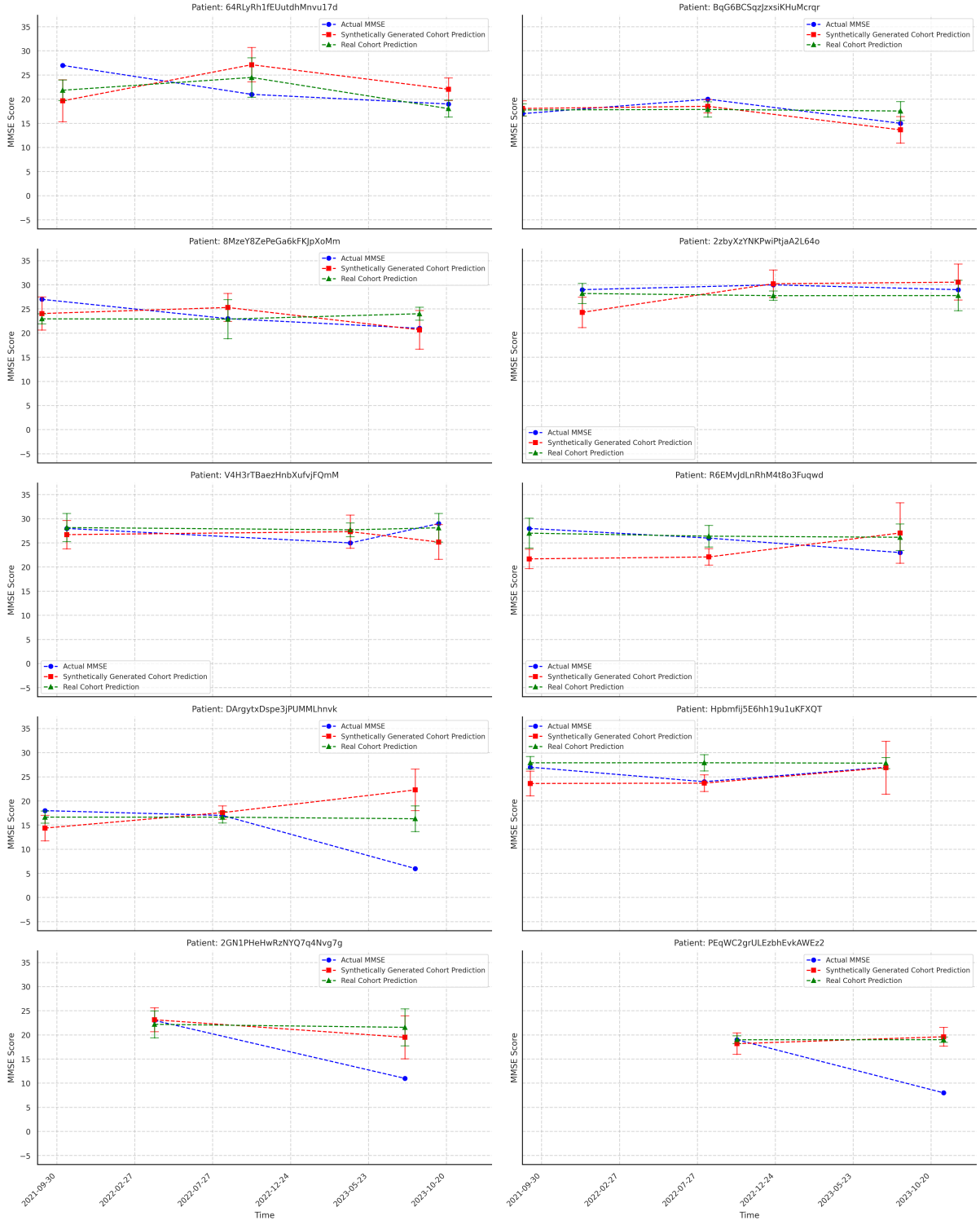


Figure 22: MMSE Prognostic Results on Test set

6.4.1 Urinary Tract Infection Risk Analysis

To evaluate the consistency between empirical and virtual raters, we performed a classification consistency assessment. The results show a Cohen Kappa coefficient of 0.626, indicating substantial agreement between the two raters. Furthermore, the percent agree-

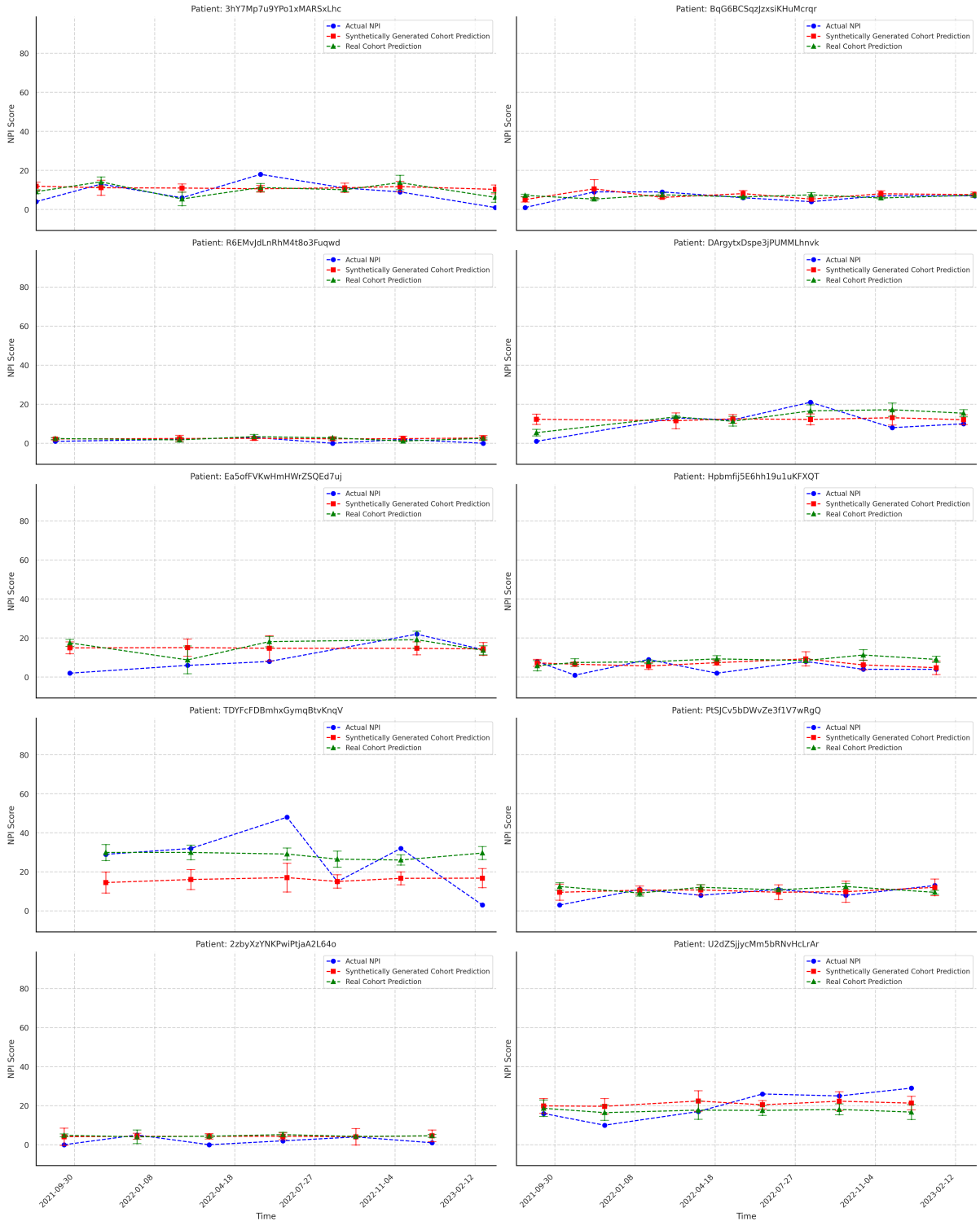


Figure 23: NPI Prognostic Results on Test set

ment between the two raters is 84.2%, further supporting a high level of concordance.

Table 7 presents the diagnostic performance metrics for both raters. The empirical rater achieves a sensitivity of 0.868 and a specificity of 0.620, resulting in an overall accuracy of 0.795. The virtual rater demonstrates superior performance, with a sensitivity of 0.909 and a specificity of 0.900, leading to an accuracy of 0.906. In particular, the

virtual rater achieves a higher positive predictive value (PPV) of 0.957 compared to 0.847 for the empirical rater, indicating a lower false positive rate.

Table 7: Diagnostic Performance Comparison for UTI Classification

Rater	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**Accuracy**
classification of Real data	0.868	0.620	0.847	0.660	0.795
classification of Synthetic data	0.909	0.900	0.957	0.804	0.906

To further assess the agreement between the empirical and virtual raters, a Bland–Altman plot was generated, as shown in Figure 24. The Bland-Altman plot suggests strong agreement between the virtual and empirical cohorts, with a small mean difference and most data points falling within the limits of agreement. This supports the validity of the virtual cohort for downstream analysis and modeling, although some minor deviations at extreme values may need further investigation.

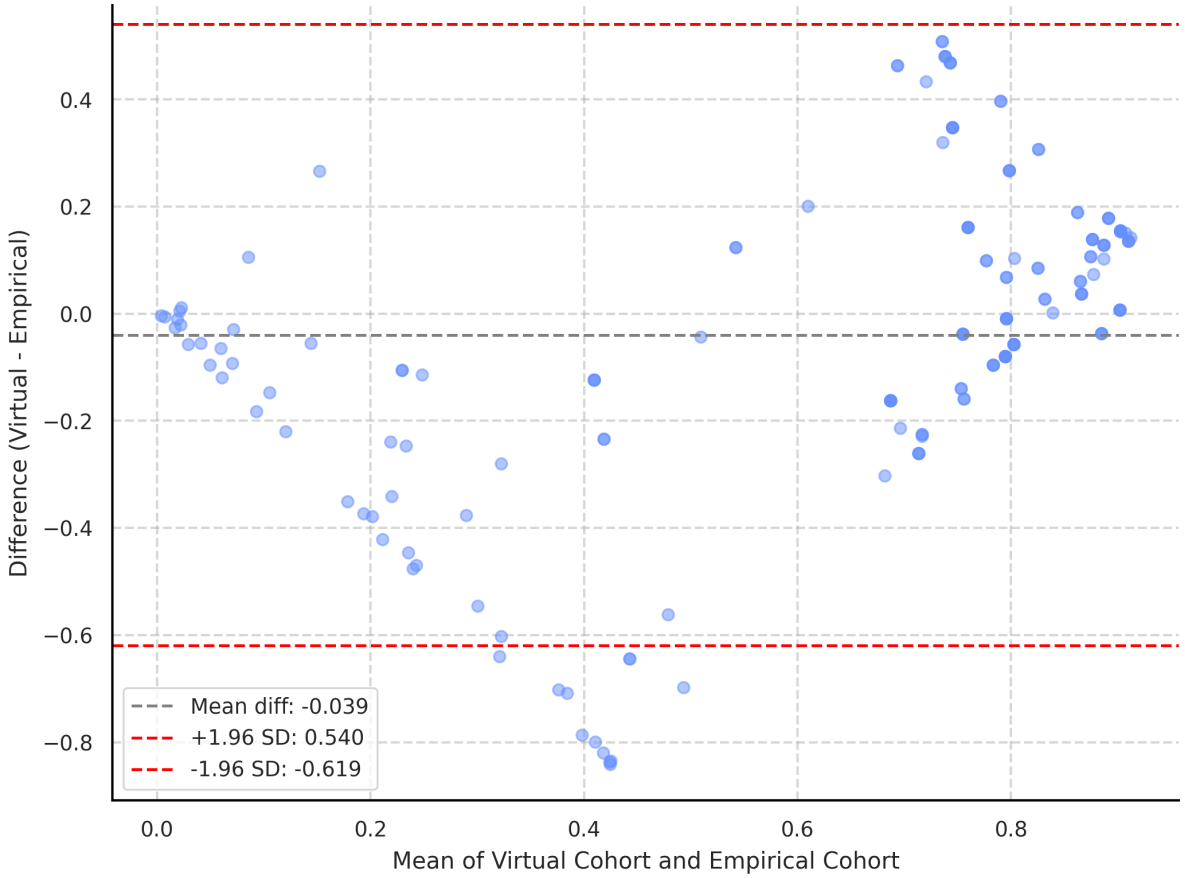


Figure 24: Bland–Altman for UTI

To quantify the reliability between raters, we computed the Intraclass Correlation Coefficient (ICC) under different assumptions, as summarized in Table 8. ICC values for single-rater (ICC1, ICC2, and ICC3) range from 0.572 to 0.576, indicating moderate reliability when individual raters are considered. When averaging across multiple raters (ICC1k, ICC2k, ICC3k), the reliability improves significantly, with ICC values

ranging from 0.728 to 0.731. These results suggest that the virtual rater achieves greater consistency when aggregating multiple ratings.

Table 8: Intraclass Correlation Coefficients (ICC) for UTI Rater Agreement

Type	Description	ICC	F	df1	df2	p-value
ICC1	Single raters absolute	0.572	3.677	170	171	1.02×10^{-16}
ICC2	Single random raters	0.573	3.720	170	170	6.61×10^{-17}
ICC3	Single fixed raters	0.576	3.720	170	170	6.61×10^{-17}
ICC1k	Average raters absolute	0.728	3.677	170	171	1.02×10^{-16}
ICC2k	Average random raters	0.729	3.720	170	170	6.61×10^{-17}
ICC3k	Average fixed raters	0.731	3.720	170	170	6.61×10^{-17}

To further interpret the ICC values, Table 9 presents the 95% confidence intervals (CI) for each ICC estimate. ICC values for single-rater have wider confidence intervals (e.g. ICC1: [0.46, 0.66]), reflecting greater variability in individual ratings. In contrast, when averaging between multiple raters, the confidence intervals become narrower (e.g. ICC3k: [0.64, 0.80]), indicating increased reliability and reduced uncertainty.

Table 9: 95% Confidence Intervals for ICC Estimates in UTI CLassification

ICC Type	95% CI
ICC1	[0.46, 0.66]
ICC2	[0.46, 0.67]
ICC3	[0.47, 0.67]
ICC1k	[0.63, 0.80]
ICC2k	[0.63, 0.80]
ICC3k	[0.64, 0.80]

Figure 25 presents the classification results for urinary tract infection (UTI) prediction in multiple patients in the test set, comparing the predictions of the empirical (real) cohort and the virtual cohort over time. Each subplot corresponds to an individual patient, where the true status of UTI is indicated by red (UTI positive) and green (UTI negative) bars at the top. The predicted UTI probabilities are visualized with the blue line for the real cohort and the purple line for the virtual cohort. The results show that the predictions for the real cohort closely align with the observed UTI status, suggesting that the model effectively captures infection episodes. Meanwhile, virtual cohort predictions generally follow similar trends but exhibit minor deviations, highlighting subtle differences in synthetic patient behavior. In some cases, the probability of UTI remains elevated despite a transition in ground truth status, possibly due to the model's sensitivity to historical patterns. Overall, the results validate that the virtual cohort can approximate the trajectories of UTI in real patients with a reasonable degree of precision, supporting its potential for future predictive modeling and clinical applications.

6.4.2 Agitation Risk Analysis

The agreement between empirical and virtual raters for the classification of agitation was evaluated using Cohen's Kappa and percent agreement. Cohen's Kappa coefficient was

found to be -0.121, indicating poor agreement, suggesting that the two rater classifications are not consistently aligned. Additionally, the percent agreement was 46.7%, which is close to random chance (50%), further emphasizing the lack of concordance.

The diagnostic performance metrics for the empirical and virtual raters are summarized in Table 10. The empirical rater achieves a sensitivity of 0.735 and an accuracy of 0.589, while the virtual rater demonstrates a lower sensitivity of 0.633 and an overall accuracy of 0.523. Both raters exhibit low specificity (0.466 and 0.431, respectively), indicating a high false positive rate. The positive predictive value (PPV) and negative predictive value (NPV) are also relatively low, suggesting that both raters struggle to reliably distinguish between agitation and nonagitation cases.

Table 10: Diagnostic Performance Comparison for Agitation Classification

Rater	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**Accuracy**
classification of Real data	0.735	0.466	0.537	0.675	0.589
classification of Synthetic data	0.633	0.431	0.484	0.581	0.523

A Bland–Altman plot was also generated for agitation, as shown in Figure 26. Although this graph indicates a high level of agreement, the other metrics do not support this conclusion.

ICC analysis provides additional information on the level of agreement between raters (Table 11). The single-rater ICC values (ICC1, ICC2, ICC3) are negative (-0.328 to -0.307), indicating poor reliability and suggesting that the variance between raters is greater than the variance within individual ratings. When averaging multiple raters (ICC1k, ICC2k, ICC3k), the ICC values drop further to -0.977 to -0.886, indicating that aggregation does not improve agreement and may even exacerbate inconsistencies.

Table 11: Intraclass Correlation Coefficients (ICC) for Agitation Rater Agreement

Type	**Description**	**ICC**	**F**	**df1**	**df2**	**p-value**
ICC1	Single raters absolute	-0.328	0.506	106	107	0.9997
ICC2	Single random raters	-0.307	0.518	106	106	0.9996
ICC3	Single fixed raters	-0.317	0.518	106	106	0.9996
ICC1k	Average raters absolute	-0.977	0.506	106	107	0.9997
ICC2k	Average random raters	-0.887	0.518	106	106	0.9996
ICC3k	Average fixed raters	-0.929	0.518	106	106	0.9996

ICC values indicate poor agreement and reliability in the agitation classification task. Generally, ICC values below 0.50 indicate poor reliability, and negative ICC values suggest that the variance between the raters exceeds the variance within the individual ratings, meaning that the two raters do not measure agitation consistently. The confidence intervals (CI95%) further confirm the lack of reliability, as they span entirely negative values, such as ICC1 [-0.49, -0.15] and ICC3k [-1.83, -0.32] (Table 12).

These results highlight significant inconsistencies between empirical and virtual raters, suggesting that the current agitation classification model may require further optimization. The low specificity and low positive predictive value (PPV) indicate a high false positive rate, which can lead to misclassification. Additionally, the negative ICC values suggest fundamental inconsistencies in rating agitation, which may be due to subjective

Table 12: 95% Confidence Intervals for ICC Estimates in Agitation Classification

ICC Type	95% CI
ICC1	[-0.49, -0.15]
ICC2	[-0.47, -0.13]
ICC3	[-0.48, -0.14]
ICC1k	[-1.90, -0.35]
ICC2k	[-1.76, -0.29]
ICC3k	[-1.83, -0.32]

differences in interpretation or limitations in the model's ability to generalize from prior knowledge.

The agitation rater analysis reveals poor agreement between empirical and virtual raters, with low classification accuracy and reliability. The ICC results further confirm that the agitation classification task is highly inconsistent and averaging multiple raters does not improve reliability.

Figure 27 presents the classification results for predicting agitation in multiple patients from the test set, comparing the predictions of the empirical (real) cohort and the virtual cohort over time. Each subplot represents an individual patient, illustrating the model's ability to identify episodes of agitation. Observations from the figure suggest that both the real cohort predictions and the virtual cohort predictions exhibit limitations in accurately capturing agitation episodes. While the real cohort predictions align with the agitation labels in some cases, there are noticeable discrepancies where the predicted probability does not match the true agitation status. This suggests that the model struggles to generalize well to the test set. The performance of the virtual cohort is markedly worse, with predicted probabilities often failing to track the real agitation labels. This suggests that synthetically generated agitation trajectories do not accurately replicate the behavioral patterns observed in real patients, resulting in poor classification results.

6.4.3 Comparison Results

Compared to the UTI classification results, the poor agreement and low reliability observed in the agitation classification appear to be highly reasonable. This discrepancy can be attributed to the manner in which prior information was incorporated into the data generation process.

UTI data was explicitly included as prior information when generating virtual cohort samples, allowing the large language model (LLM) to leverage its preexisting knowledge and enhance classification performance. For example, the LLM may have learned to associate specific movement patterns with UTI cases, making it easier for virtual cohort samples to be correctly classified. This explains why the virtual rater achieved high diagnostic performance in the UTI classification task.

In contrast, agitation information was not provided as a prior knowledge of the model. Since the virtual cohort was generated without any information related to agitation, the model had no structured way to encode agitation patterns into the generated samples. As a result, the classification of agitation is poor, leading to a near-chance level agreement between the empirical and virtual raters.

However, the interpretation of the agitation prediction on the virtual cohort warrants a critical discussion regarding data leakage. The observed results (Kappa = -0.121) sparked debate during the evaluation: while the poor performance was expected due to the withheld labels, the specific negative correlation suggests that the model behavior is not purely random. This raises the concern that the movement embeddings themselves—which were used to generate the virtual cohort—might still retain 'leaked' structural information associated with agitation events. If the embeddings perfectly decoupled movement from agitation, the result should theoretically approach zero. The deviation implies that even without explicit labels, the foundation model may generate artifacts or inverse patterns derived from the underlying behavioral data, indicating that complete isolation of clinical variables in high-dimensional healthcare data remains a significant challenge.

6.5 Adversarial Evaluation of Synthetic Data Quality

To quantitatively assess the quality and realism of the generated virtual cohort trajectories, we designed an adversarial evaluation framework. The core principle of this approach is to train a discriminator model tasked with distinguishing between real patient trajectories and synthetically generated ones. The performance of this discriminator serves as an inverse metric of the quality of the synthetic data: a lower discriminator accuracy implies that the generated data are more realistic and statistically indistinguishable from real data, indicating higher quality of generation.

The discriminator is designed to serve as a powerful feature extractor for trajectory sequences. The key design choices and hyperparameters are as follows:

- **Base Model**: We use a pre-trained `distilbert-base-uncased` model as the foundational feature extractor. To maintain its robust, pre-trained knowledge of sequence representation, its parameters were frozen during training.
- **Classification Head**: A lightweight classification head was added on top of the frozen base model. This head consists of a dropout layer (rate = 0.3) for regularization. This single linear layer maps features to a logit, and a sigmoid activation function to produce a final probability score.
- **Tokenizer Max Length**: Input sequences were padded or truncated to a maximum length of $L_{max} = 64$.
- **Training Parameters**: The model was trained for $E = 20$ epochs with a batch size of $B = 4$. We used the Adam optimizer with a learning rate of $\eta = 2 \times 10^{-5}$ and Binary Cross-Entropy (`BCELoss`) as the loss function.

6.5.1 Results and Interpretation

After training, the adversarial discriminator achieved an accuracy of **72.87%** on the holdout test set. This result indicates that the discriminator possesses a discernible, though not overwhelming, ability to distinguish between the real and synthetic movement trajectories. An accuracy of more than 70% suggests that, while the generative model has successfully captured some of the significant patterns and high-level characteristics of the empirical data, there are still remaining statistical differences or artifacts that the discriminator can identify. This provides a clear, quantitative signal that the synthetic

data are reasonably realistic but have significant room for improvement to more closely mimic the complex, nuanced dynamics of real-world patient behavior.

6.5.2 Limitations regarding Privacy-Preserving Properties

While the adversarial evaluation above focuses on the realism of the data, the examiners have rightly raised concerns regarding the 'privacy-preserving' nature of the virtual cohort. Since the RAG framework explicitly retrieves real patient data to condition the generation process, the resulting virtual patients are likely to be very close to real patient data in the high-dimensional feature space. Therefore, the virtual patients may in fact be disclosing information about real patients.

We acknowledge that a proper assessment of the privacy-preserving properties would require a more thorough analysis involving well-defined threat models. Specifically, future work must subject the model to attacks such as membership inference attacks (determining if a specific individual was used in the training/retrieval set) and model inversion attacks (attempting to reconstruct sensitive features of real patients from the virtual outputs). Until such rigorous stress-testing is performed, the privacy protection offered by this RAG-based approach should be viewed as a method of obfuscation rather than cryptographic security.

6.6 Sensitivity Analysis of Input Time Windows

We also conducted a sensitivity analysis (parameter sweep) to investigate the impact of using different time-lengths of PLWD mobility trajectory data as input. Unlike the component ablation study which evaluated architectural choices, this analysis focuses on optimizing the temporal hyperparameters. In previous experiments, we used the preceding 30 days as input, a choice that clinical experts endorsed. However, clinicians were particularly interested in whether shorter time frames could yield reliable predictions. Additionally, we aimed to examine whether variations in input length would significantly affect model performance.

Table 13: MAE Sensitivity Analysis Results over Different Time Windows (with 95% CI), Grouped by Cohort and Downstream Tasks

Cohort	Downstream Task	7 days	14 days	30 days	60 days	90 days	180 days
REAL	MMSE	4.1004 (3.3940, 4.8068)	3.5775 (2.9983, 4.1568)	4.2459 (3.4774, 5.0144)	4.2939 (3.6120, 4.9757)	3.8094 (3.0400, 4.5789)	4.4049 (3.6864, 5.1234)
	ADAS-COG	11.5260 (10.3043, 12.7477)	11.9230 (10.9302, 12.9159)	11.8717 (10.8215, 12.9219)	11.4882 (10.3163, 12.6601)	11.1194 (10.1239, 12.1148)	11.4764 (10.1240, 12.8288)
	NPI	8.8843 (7.5661, 10.2026)	8.4658 (7.0987, 9.8329)	8.6685 (7.3236, 10.0134)	8.9952 (7.5302, 10.4602)	9.2876 (7.7817, 10.7936)	8.6430 (6.9381, 10.3479)
VIRTUAL	MMSE	4.3248 (3.5974, 5.0521)	3.9177 (3.2944, 4.5410)	4.2879 (3.4776, 5.0982)	4.5623 (3.7780, 5.3466)	4.0275 (3.2690, 4.7861)	4.8597 (4.0268, 5.6927)
	ADAS-COG	12.0687 (10.7930, 13.3444)	12.2057 (11.1573, 13.2541)	12.2449 (11.2045, 13.2853)	12.0556 (10.8304, 13.2808)	11.7725 (10.7254, 12.8196)	11.5166 (10.0547, 12.9784)
	NPI	9.0870 (7.8684, 10.3055)	9.0224 (7.6025, 10.4424)	8.9213 (7.6770, 10.1656)	9.0011 (7.4860, 10.5161)	9.5175 (8.0407, 10.9943)	9.0282 (7.3399, 10.7165)

We also conducted an ablation study to investigate the impact of using different time-lengths of PLWD mobility trajectory data as input. In previous experiments, we used the preceding 30 days as input, a choice that clinical experts endorsed. However, clinicians were particularly interested in whether shorter time frames could yield reliable predictions. Additionally, we aimed to examine whether variations in input length would significantly affect model performance, with a particular focus on the predictive outcomes for the generated virtual cohort. Given that the virtual cohort's input is derived from pre-trained seven-day mobility data, it is particularly meaningful to assess whether its predictive performance remains comparable to that of the real dataset over longer time windows.

As indicated by the highlighted results in Table 13, half of the optimal performance values were obtained with a 14-day input duration, which aligns well with our expectations. In contrast, the best-performing input lengths for the virtual cohort were more dispersed. We hypothesize that this can be attributed to the presence of long-term features, such as HADS scores, age, sex, and type of dementia, in the generated data, which could allow the model to maintain strong predictive performance even when provided with longer input time windows.

In Figure 28, it can be seen that, for each individual cognitive assessment metric, the predictive results for the virtual cohort were consistently slightly worse than those for the real data set. This suggests that while our generative pipeline is capable of capturing the deep representations of patient behavioral patterns, it has not yet fully achieved the level of downstream task performance that the real dataset provides.

7 Conclusions and Future Work

Our initial results demonstrate that the latent state vectors derived from PLWD's daily activity patterns can effectively capture behavioral dynamics within our framework. Based on these findings, we then introduced a novel framework leveraging post-training recovery-agmented generation (RAG) to construct structured virtual cohorts for the anonymization and augmentation of healthcare temporal data. By integrating contrastive learning-based embeddings with generative large language models (LLM), our approach enables the creation of high-quality synthetic datasets that preserve the statistical and predictive characteristics of real-world clinical data. Through extensive evaluations using unsupervised metrics and downstream tasks, including urinary tract infection prediction, cognitive evaluations, and agitation detection, we demonstrated that virtual cohorts generated by our method achieve performance comparable or superior to real datasets while offering a potential pathway for data sharing. However, we recognize that the proximity of generated data to real patient records—inherent to the RAG mechanism—poses valid privacy concerns. Future work will explore further optimization of retrieval strategies and augmentation techniques to enhance the fidelity and diversity of generated data. Furthermore, integrating domain-specific constraints and expert feedback could refine the realism and applicability of synthetic data sets in more specialized clinical settings.

From a clinical perspective, a key challenge is aligning the outputs of AI models with actionable insights for healthcare providers. The proposed low-rank state representations must closely correlate with clinically significant metrics, such as MMSE and ADAS-Cog scores, to support diagnosis and intervention planning. However, the limited size and scope of our dataset, which includes only 134 participants with complete data from 50 patients, presents a significant limitation. This small sample size may not fully capture the diversity of movement and behavioral patterns observed in the broader population of people living with dementia, potentially compromising the model's generalizability to other settings. Furthermore, variability in patient behavior due to external factors, such as seasonal changes or caregiver interactions, requires models that can differentiate between pathological changes and normal variations.

Future work will focus on optimizing the framework to enhance the fidelity and diversity of the generated data. This includes further refining retrieval strategies and augmentation techniques within the RAG model and integrating domain-specific constraints and expert feedback to improve the realism and applicability of the synthetic datasets in more specialized clinical settings. Addressing the current limitations requires an interdisciplinary approach, incorporating expertise from AI researchers, clinicians, ethicists, and regulatory bodies to develop robust and clinically impactful monitoring solutions. Critically, expanding the dataset to include a more diverse and representative sample of patients will be essential to improve the reliability, applicability, and equity of the model's predictions. This will help ensure that the developed healthcare monitoring solutions are both effective and fair.

In conclusion, the research provides a novel approach to developing foundation models from in-home monitoring data. This complements conventional approaches to analyse the data to detect changes in behaviour and activity patterns and also develop predictive models for health risks such as UTIs and agitation in dementia. We have demonstrated that using a semi-supervised model can provide a flexible and generalisable approach to analyse large-scale in-home monitoring data in dementia.

References

[1] Yushan Huang, Yuchen Zhao, Alexander Capstick, Francesca Palermo, Hamed Haddadi, and Payam Barnaghi. Analyzing entropy features in time-series data for pattern recognition in neurological conditions. *Artificial Intelligence in Medicine*, 150:102821, 2024.

[2] Dokyun Kim, Sukhyun Cho, Heewoong Chae, Jonghun Park, and Jaeseok Huh. Semi-supervised contrastive learning with decomposition-based data augmentation for time series classification. *Intelligent Data Analysis*, Preprint(Preprint):1–25, January 2024. Publisher: IOS Press.

[3] Haoyi Fan, Fengbin Zhang, Ruidong Wang, Xunhua Huang, and Zuoyong Li. Semi-Supervised Time Series Classification by Temporal Relation Prediction. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3545–3549, June 2021. ISSN: 2379-190X.

[4] Jean-Yves Franceschi, Aymeric Dieuleveut, and Martin Jaggi. Unsupervised Scalable Representation Learning for Multivariate Time Series. *ArXiv*, January 2019.

[5] Sangho Lee, Wonjoon Kim, and Youngdoo Son. Spatio-Temporal Consistency for Multivariate Time-Series Representation Learning. *IEEE Access*, 12:30962–30975, 2024. Conference Name: IEEE Access.

[6] Xinrui Lyu, Matthias Hueser, Stephanie L. Hyland, George Zerveas, and Gunnar Raetsch. Improving Clinical Predictions through Unsupervised Time Series Representation Learning, December 2018. arXiv:1812.00490 [cs, stat].

[7] Chengyang Ye and Qiang Ma. LBP4MTS: Local Binary Pattern-Based Unsupervised Representation Learning of Multivariate Time Series. *IEEE Access*, 11:118595–118605, 2023. Conference Name: IEEE Access.

[8] Andrew Hill, Russell Bowler, Katerina Kechris, and Farnoush Banaei-Kashani. Semi-supervised Embedding for Scalable and Accurate Time Series Clustering. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 942–951, December 2022.

[9] Vassiliki Koufi, Flora Malamateniou, and George Vassilacopoulos. A Big Data-driven Model for the Optimization of Healthcare Processes. In *Digital Healthcare Empowering Europeans*, pages 697–701. IOS Press, 2015.

[10] Md Nagib Mahfuz Sunny, Mohammad Balayet Hossain Saki, Abdullah Al Nahian, Syed Walid Ahmed, Md Newaz Shorif, Jennet Atayeva, and Syed Wajeeh Abbs Rizvi. Optimizing Healthcare Outcomes through Data-Driven Predictive Modeling. *Journal of Intelligent Learning Systems and Applications*, 16(04):384–402, 2024.

[11] Ciro Mennella, Umberto Maniscalco, Giuseppe De Pietro, and Massimo Esposito. Ethical and regulatory challenges of AI technologies in healthcare: A narrative review. *Heliyon*, 10(4), February 2024. Publisher: Elsevier.

[12] Pouyan Esmaeilzadeh. Challenges and strategies for wide-scale artificial intelligence (AI) deployment in healthcare practices: A perspective for healthcare organizations. *Artificial Intelligence in Medicine*, 151:102861, May 2024.

[13] Arenike Patricia Adekugbe and Chidera Victoria Ibeh. ADVANCING HEALTHCARE DATA SOLUTIONS: COMPARATIVE ANALYSIS OF BUSINESS AND RESEARCH MODELS IN THE U.S. *International Medical Science Research Journal*, 4(4):373–390, April 2024. Number: 4.

[14] Muhammad Usman Hadi, Qasem Al Tashi, Abbas Shah, Rizwan Qureshi, Amgad Muneer, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, and Mubarak Shah. Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects, August 2024.

[15] Andrzej Maćkiewicz and Waldemar Ratajczak. Principal components analysis (PCA). *Computers & Geosciences*, 19(3):303–342, March 1993.

[16] M. L. Menéndez, J. A. Pardo, L. Pardo, and M. C. Pardo. The Jensen-Shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318, March 1997.

[17] Artem Obukhov and Mikhail Krasnyanskiy. Quality Assessment Method for GAN Based on Modified Metrics Inception Score and Fréchet Inception Distance. In Radek Silhavy, Petr Silhavy, and Zdenka Prokopova, editors, *Software Engineering Perspectives in Intelligent Systems*, pages 102–114, Cham, 2020. Springer International Publishing.

[18] Jiska Cohen-mansfield, Marcia S. Marx, and Alvin S. Rosenthal. A Description of Agitation in a Nursing Home. *Journal of Gerontology*, 44(3):M77–M84, May 1989.

[19] Mohd Faizan Rizvi. ARIMA Model Time Series Forecasting. *International Journal for Research in Applied Science and Engineering Technology*, 12(5):3782–3785, May 2024.

[20] Vaia I. Kontopoulou, Athanasios D. Panagopoulos, Ioannis Kakkos, and George K. Matsopoulos. A Review of ARIMA vs. Machine Learning Approaches for Time Series Forecasting in Data Driven Networks. *Future Internet*, 15(8):255, August 2023. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.

[21] Ricardo P. Masini, Marcelo C. Medeiros, and Eduardo F. Mendes. Machine learning advances for time series forecasting. *Journal of Economic Surveys*, 37(1):76–111, 2023. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/joes.12429.

[22] Maryem Rhanoui, Siham Yousfi, Mounia Mikram, and Hajar Merizak. Forecasting financial budget time series: ARIMA random walk vs LSTM neural network. *IAES International Journal of Artificial Intelligence (IJ-AI)*, 8(4):317–327, December 2019. Number: 4.

[23] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon

Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners, May 2020.

[24] Yuxuan Bian, Xuan Ju, Jiangtong Li, Zhijian Xu, Dawei Cheng, and Qiang Xu. Multi-Patch Prediction: Adapting LLMs for Time Series Representation Learning, February 2024.

[25] Yong Liu, Guo Qin, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. AutoTimes: Autoregressive Time Series Forecasters via Large Language Models, February 2024.

[26] Alexander Capstick, Tianyu Cui, Yu Chen, and Payam Barnaghi. Representation learning of daily movement data using text encoders. *ICLR 2024 Workshop Time Series for Health*, 2024.

[27] Lawrence Page, Sergey Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking : Bringing Order to the Web. November 1999.

[28] Gábor Iván and Vince Grolmusz. When the Web meets the cell: using personalized PageRank for analyzing protein interaction networks. *Bioinformatics*, 27(3):405–407, February 2011.

[29] Dániel Bánky, Gábor Iván, and Vince Grolmusz. Equal Opportunity for Low-Degree Network Nodes: A PageRank-Based Method for Protein Target Identification in Metabolic Graphs. *PLOS ONE*, 8(1):e54204, January 2013. Publisher: Public Library of Science.

[30] Atiye Sadat Hashemi, Kobra Etminani, Amira Soliman, Omar Hamed, and Jens Lundström. Time-series Anonymization of Tabular Health Data using Generative Adversarial Network. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, June 2023. ISSN: 2161-4407.

[31] Acar Tamersoy, Grigorios Loukides, Mehmet Ercan Nergiz, Yucel Saygin, and Bradley Malin. Anonymization of Longitudinal Electronic Medical Records. *IEEE Transactions on Information Technology in Biomedicine*, 16(3):413–423, May 2012. Conference Name: IEEE Transactions on Information Technology in Biomedicine.

[32] Morvarid Sehatkar and Stan Matwin. HALT: Hybrid anonymization of longitudinal transactions. In *2013 Eleventh Annual Conference on Privacy, Security and Trust*, pages 127–134, July 2013.

[33] Ziqi Zhang, Chao Yan, Thomas A Lasko, Jimeng Sun, and Bradley A Malin. SynTEG: a framework for temporal structured electronic health data simulation. *Journal of the American Medical Informatics Association*, 28(3):596–604, March 2021.

[34] Chufan Gao, Mandis Beigi, Afrah Shafquat, Jacob Aptekar, and Jimeng Sun. TrialSynth: Generation of Synthetic Sequential Clinical Trial Data, September 2024. arXiv:2409.07089.

[35] Yuan Zhong, Xiaochen Wang, Jiaqi Wang, Xiaokun Zhang, Yaqing Wang, Mengdi Huai, Cao Xiao, and Fenglong Ma. Synthesizing Multimodal Electronic Health Records via Predictive Diffusion Models, June 2024. arXiv:2406.13942.

[36] Brandon Theodorou, Cao Xiao, and Jimeng Sun. Synthesize Extremely High-dimensional Longitudinal Electronic Health Records via Hierarchical Autoregressive Language Model, March 2023. ISSN: 2693-5015.

[37] Simon Bing, Andrea Dittadi, Stefan Bauer, and Patrick Schwab. Conditional generation of medical time series for extrapolation to underrepresented populations. *PLOS Digital Health*, 1(7):e0000074, July 2022. Publisher: Public Library of Science.

[38] Alexander Ylnner Choquenaira Florez, Lucas Scabora, Sihem Amer-Yahia, and José Fernando Rodrigues Júnior. Augmentation Techniques for Sequential Clinical Data to Improve Deep Learning Prediction Techniques. In *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 597–602, July 2020. ISSN: 2372-9198.

[39] Hyunwoo Sohn, Kyungjin Park, Baekkwan Park, and Min Chi. Multi-TA: Multilevel Temporal Augmentation for Robust Septic Shock Early Prediction. volume 7, pages 6035–6043, August 2024. ISSN: 1045-0823.

[40] Shenglai Zeng, Jiankun Zhang, Pengfei He, Jie Ren, Tianqi Zheng, Hanqing Lu, Han Xu, Hui Liu, Yue Xing, and Jiliang Tang. Mitigating the Privacy Issues in Retrieval-Augmented Generation (RAG) via Pure Synthetic Data, June 2024. arXiv:2406.14773.

[41] Rui Yang, Yilin Ning, Emilia Keppo, Mingxuan Liu, Chuan Hong, Danielle S. Bitterman, Jasmine Chiat Ling Ong, Daniel Shu Wei Ting, and Nan Liu. Retrieval-Augmented Generation for Generative Artificial Intelligence in Medicine, June 2024. arXiv:2406.12449.

[42] Hajra Murtaza, Musharif Ahmed, Naurin Farooq Khan, Ghulam Murtaza, Saad Zafar, and Ambreen Bano. Synthetic data generation: State of the art in health care domain. *Computer Science Review*, 48:100546, May 2023.

[43] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-Pack: Packaged Resources To Advance General Chinese Embedding, 2023. _eprint: 2309.07597.

[44] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization, January 2019. arXiv:1711.05101 [cs, math].

[45] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. MTEB: Massive Text Embedding Benchmark, March 2023. arXiv:2210.07316 [cs].

[46] Laurens van der Maaten and Geoffrey Hinton. Viualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, November 2008.

[47] Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison,

Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, CK Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Michael Suo, Phil Tillet, Eikan Wang, Xiaodong Wang, William Wen, Shunting Zhang, Xu Zhao, Keren Zhou, Richard Zou, Ajit Mathews, Gregory Chanan, Peng Wu, and Soumith Chintala. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation, April 2024. Publication Title: 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24) original-date: 2016-08-13T05:26:41Z.

[48] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Perric Cistac, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-Art Natural Language Processing. pages 38–45. Association for Computational Linguistics, October 2020.

[49] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and others. Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[50] Charles R. Harris, K. Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585:357–362, 2020.

[51] Wes McKinney and others. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX, 2010.

Appendix A Appendix

A.1 Participants Demographics and Clinical Characteristics

A.1.1 Demographics

A total of **134** patients were included in the dataset. The age of the patients ranged from **54** to **100** years. The gender distribution comprised **67** females and **67** males. Regarding living situation, **55** patients lived alone, while **61** patients lived with others.

The primary diagnoses covered various conditions, including:

- Alzheimer's Disease: 82
- Mild Cognitive Impairment (MCI): 20
- Mixed Alzheimer's and Vascular Dementia: 15
- Vascular Dementia: 12
- Other diagnoses including frailty, frontotemporal dementia, stroke, and Lewy Body Dementia.

A.1.2 Comorbidities

The cohort presented diverse comorbidities, including cardiovascular diseases, endocrine disorders, musculoskeletal conditions, and malignancies. The most prevalent conditions included:

- Cardiovascular diseases: up to 74 patients affected
- Endocrine and metabolic disorders: 33 patients affected
- Musculoskeletal conditions: 54 patients affected
- Psychiatric disorders: 21 patients affected
- Malignancies: at least 10 patients diagnosed

A.1.3 In-Home Movement Trajectories

PLWD's daily movement was documented over the period from **2021-07-06** to **2024-01-15** within their homes was tracked across predefined locations:

- Bedroom, Bathroom, Kitchen, Lounge, Hallway, Sleep, Wake, None

A.1.4 Agitation, Mood, and UTI Status

Mood and agitation data were collected for **67** patients, spanning from **2021-07-22** to **2023-03-15**.

UTI status was documented over the period from **2021-07-06** to **2024-01-15**, with the following distribution:

- UTI Positive Days: 18536

- UTI Negative Days: 47560
- Missing or Unknown: 0

The HADS Anxiety scores ranged from **0.0** to **19.0**.

A.2 Distribution of Cognitive and Behavioral Assessments

Longitudinal cognitive and behavioral assessments were performed, covering the period from **2022-08-03** to **2024-04-02**.

A.2.1 Temporal Distribution of Clinical Records

The temporal distribution of records in the dataset used in this study is shown in Figure 8.

A.3 Datasets and Code Availability

The IPython notebooks used to develop this framework will be released following review. However, due to data sensitivity, the dataset and notebooks used for data visualization will not be publicly available. All experiments were conducted using Python 3.11.5, with the following libraries: Torch 2.4.0 [47], Transformers 4.44.2 [48], Sentence-Transformers 2.7.0 [49], Scikit-Learn 1.3.2 [49], NumPy 1.26.4 [50], SciPy 1.13.1, and Pandas 2.1.4 [51].

A.4 Location time series and distribution visualization for each cluster

Following extraction of location embeddings, we identified five groups based on the highest silhouette score, achieved through k-means clustering. The resulting distribution presents the temporal clustering of each location for an individual participant, as is shown in Figure 29, 30, 31, 32, 33, along with the cumulative distribution of each location across all participants in Figure 34, 36 and 35.

A.5 T-SNE plots for individual participants in test set

Figures 37 and 38 illustrate individual movement patterns within the embedded space, depicting their locations and timespans. Each data point was collected between July 31, 2023, and January 31, 2024.

A.6 Relations between most similar and dissimilar pariticipants

This section summarizes the analysis of patient characteristics based on similarity metrics calculated from PageRank state embeddings. The aim is to compare characteristics between selected participants, in this case, using all 50 participants, their three most similar counterparts, and their least similar counterparts. The rate of change in MMSE and ADAS-Cog scores ($\Delta MMSE$ and $\Delta ADAS - Cog$) was calculated by subtracting the participants' scores from one year prior from their current scores within the selected time period and normalizing the result by the time difference.

Table 14, 15, 16, 17 indicates that there are various degrees of differences between selected patients and their most and least similar counterparts. The p-values suggest

Table 14: P-values for characteristics comparison with most similar patients.

Feature	**P-Value** (n=50)
MMSE	0.9978
ADAS-Cog	0.9788
HADS - Depression Score	0.8171
HADS - Anxiety score	0.8239
Age	0.8860
$\Delta\ MMSE$	0.3773
$\Delta\ ADAS-Cog$	0.5628

Table 15: P-values for characteristics comparison with least similar patients.

Feature	**P-Value** (n=50)
MMSE	0.8969
ADAS-Cog	0.8696
HADS - Depression Score	0.2560
HADS - Anxiety score	**0.0069**
Age	**4.5934e-05**
$\Delta\ MMSE$	**0.0224**
$\Delta\ ADAS-Cog$	0.4426

Table 16: Effect sizes for comparisons with most similar patients.

Feature	**Effect Size (Cohen's d)**
MMSE	-1.3979
ADAS-Cog	2.3360
HADS - Depression Score	1.2452
HADS - Anxiety score	2.7198
Age	0.8310
$\Delta\ MMSE$	-0.0999
$\Delta\ ADAS-Cog$	-0.7578

Table 17: Effect sizes for comparisons with least similar patients.

Feature	**Effect Size (Cohen's d)**
MMSE	-0.9228
ADAS-Cog	3.1013
HADS - Depression Score	-2.9330
HADS - Anxiety score	-6.6526
Age	-24.8634
$\Delta\ MMSE$	-23.0227
$\Delta\ ADAS-Cog$	5.7778

that for most features, there are no statistically significant differences between the selected patients and their most similar or least similar counterparts, except for the HADS-Anxiety score, Age, and change in Adas-cog scores in the least similar group. The effect sizes provide insight into the magnitude of differences, where large effect sizes are observed in some features such as HADS - Anxiety score Age and change in Adas-Cog scores.

A.7 Clinical Significance of Deep Vector States

Through detailed analysis of the clinical relevance of deep vector states, their clustering, and patient behavior similarity, we have discovered that these deep but compact vectors encapsulate substantial information about patient behavior patterns. Each state represents a distinct cognitive or behavioral profile, with the intensity of each state reflecting specific cognitive patterns. Our objective now is to explore the potential clinical applications of these states by using them as features in machine learning models to predict MMSE and ADAS-Cog scores. These models will be trained by combining the vector states with baseline models and additional patient characteristics to assess their predictive performance.

The features included in the analysis are categorized as follows:

- **Characteristics: Clinical assessment characteristics**: This includes traditional clinical scores such as the *HADS - Depression Score*, *HADS - Anxiety Score*, clinical cognitive scores such as *MMSE Score, ADAS-Cog Score(while predicting changes of these scores)* and *Age*.

- **Characteristics: Demographic and diagnostic characteristics**: Non-numeric features like *Gender*, *Living Status (Alone or with others)*, and *clinical diagnosis (type of dementia)*.

- **Latent State Features**: The five latent states derived from deep vector representations (*state1*, *state2*, *state3*, *state4*, *state5*) which capture essential cognitive and behavioral dynamics.

- **Baseline Features**: Aggregated time series characteristics, including daily mean and variance of movement in various home locations such as *Wake Count*, *Lounge Count*, *Kitchen Count*, *Bedroom Count*, *Bathroom Count*, *Hallway Count*, and *Sleep Count*. Each location contributes two variables: the mean and variance of visit counts.

- **Frequency-Based Baseline:** This approach calculates the frequency of occurrence of various locations of activity for each participant, normalizing the count data each day to represent the percentage of time spent in different locations. For each location (*, for example, bedroom, kitchen, lounge*), statistical characteristics such as mean and variance are calculated for these normalized frequencies. These features serve as input to predictive models.

- **Random Word Baseline:** In this method, each location or activity state is assigned to a random numeric value. These random numeric embeddings replace meaningful semantic information and are used as features for model training. This baseline provides a reference point to evaluate the significance of structured embeddings or semantic features derived from location sequences.

- **Outcome Features**: The target variables for prediction, which are changes in clinical cognitive scores such as *MMSE Score,* Δ*MMSE, ADAS-Cog Score* and Δ*ADAS-Cog*.

The baseline model was constructed to align closely with the dimensionality of the latent state features; otherwise, the scale of the baseline characteristics of movement data for a single participant would be excessively large (7,776,000) and impossible to use. This baseline feature set aggregates time series data by taking the mean and variance of daily activity at different locations within the home during the testing period. When predicting cognitive changes over time, the model could incorporate current values to enhance the prediction. Note that the length of the time series input is six months in this case.

A.7.1 Model Parameters

To analyze the performance of the models, we employed three algorithms: XGBoost, LightGBM, and Support Vector Machine (SVM). The specific configurations of each model are as follows:

- **XGBoost**: Key parameters include:
  - *n_estimators*: 100, specifying the number of boosting rounds.
  - *learning_rate*: 0.1, controlling the step size shrinkage to prevent overfitting.
  - *max_depth*: 6, setting the maximum depth of each tree.
  - *random_state*: 0, ensuring reproducibility.
- **LightGBM**: Configured with the following parameters:
  - *n_estimators*: 100, specifying the number of boosting iterations.
  - *learning_rate*: 0.1, controlling the contribution of each tree.
  - *max_depth*: -1, allowing the model to grow trees without a maximum depth constraint.
  - *random_state*: 0, ensuring consistent results.
- **Support Vector Machine (SVM)**: Configured as follows:
  - *kernel*: 'rbf', employing a radial basis function for non-linear decision boundaries.
  - C: 1.0, setting the regularization parameter.
  - *epsilon*: 0.1, defining the error tolerance in the prediction.

The machine learning models were trained using data collected from 50 participants. Various sets of characteristics were explored, including baseline activity characteristics, clinical characteristics, and derived latent states, as well as combinations thereof. For preprocessing, numerical features were standardized, while categorical features were processed with the most frequent imputation strategy, followed by one-hot encoding.

The modeling approach included XGBoost, LightGBM, and SVM, which were selected for their ability to capture both linear and nonlinear relationships. Leave-One-Out Cross-Validation (LOOCV) was employed to assess the models' performance, given the small

sample size, ensuring the robustness and generalizability of the results. During each fold, the data was split into a single test sample and the remaining samples for training. Model evaluation focused on two primary metrics: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), calculated separately for MMSE and ADAS-Cog score predictions.

A.8 LSTM Model for Multivariate Time Series Prediction

Input: The input to the LSTM model consists of aggregated weekly time series features derived from patient activity data. Each characteristic represents the count of occurrences for various activity locations (*for example, bedroom, kitchen, lounge*) within a week. These time series data are standardized using `StandardScaler` to normalize feature values on different scales.

Model Architecture:

- **LSTM Layer:** A single LSTM layer with 64 units is used to capture temporal dependencies in the input data. This layer outputs a fixed-length feature vector summarizing the temporal patterns.
- **Dropout Layer:** A dropout layer with a rate of 0.2 is added to prevent overfitting by randomly setting a fraction of input units to zero during training.
- **Dense Layers:** Two fully connected dense layers are used:
  - A hidden dense layer with 64 units and ReLU activation.
  - An output dense layer with a size matching the number of target variables (*e.g., MMSE, ADAS-COG scores and their differences*).

Training Strategy:

- **Loss Function:** Mean Squared Error (MSE) is used as the loss function for training.
- **Optimizer:** The Adam optimizer is employed for efficient gradient descent.
- **Validation:** 10-fold cross-validation is applied to evaluate the model's performance. In each fold, the data is split into training and testing sets.
- **Hyperparameters:** The model is trained for 50 epochs with a batch size of 32 in each fold.

The performance metrics in Table 18 summarize the predictive accuracy of various models to forecast annual changes in the ADAS-Cog and MMSE scores (ΔADASCOG and ΔMMSE, respectively).

For the **LightGBM** model, the lowest prediction errors for ADAS-Cog were achieved using the feature set of *Random Word + Characteristics*, with an $\text{MAE}_{\Delta\text{ADASCOG}}$ of **4.95** (95% CI: 3.58, 6.42). For MMSE, the lowest prediction errors were observed with the *State + Characteristics* feature set, resulting in an $\text{MAE}_{\Delta\text{MMSE}}$ of **2.24** (95% CI: 1.67, 2.89). These findings suggest that focusing on well-selected clinical features and random word embeddings or combining state and clinical characteristics can improve predictive accuracy for both metrics.

Table 18: Performance metrics for predicting changes of ADAS-Cog and MMSE with 95% confidence intervals (CI). The values in parentheses represent the lower and upper bounds of the 95% CI. The metrics represent the predicted annual change (Δ) for MMSE and ADAS-Cog.

Model	**Metric** Feature Set	**MAE$_{\Delta ADASCOG}$**	**MAE$_{\Delta MMSE}$**	**RMSE$_{\Delta ADASCOG}$**	**RMSE$_{\Delta MMSE}$**
LightGBM	All Features	5.60 (4.24, 7.14)	2.37 (1.78, 3.04)	5.64 (4.20, 7.05)	2.34 (1.73, 2.97)
	Baseline	6.67 (4.84, 8.70)	2.57 (1.83, 3.37)	6.76 (5.05, 8.72)	2.57 (1.83, 3.37)
	Characteristics	**4.96 (3.59, 6.50)**	2.26 (1.72, 2.90)	**4.99 (3.67, 6.58)**	2.24 (1.67, 2.89)
	Proportion Baseline	6.12 (4.61, 7.89)	2.65 (2.04, 3.38)	6.14 (4.57, 7.87)	2.64 (2.04, 3.31)
	Random Word	5.69 (4.27, 7.47)	2.75 (2.10, 3.47)	5.69 (4.26, 7.36)	2.74 (2.13, 3.46)
	State	6.00 (4.41, 7.75)	2.70 (2.04, 3.37)	6.06 (4.56, 7.83)	2.69 (2.06, 3.42)
SVM	All Features	6.29 (4.77, 8.09)	3.60 (2.98, 4.29)	6.27 (4.76, 8.21)	3.59 (2.92, 4.35)
	Baseline	6.26 (4.68, 8.03)	3.58 (2.97, 4.26)	6.27 (4.75, 8.04)	3.62 (2.99, 4.37)
	Characteristics	6.19 (4.62, 7.96)	3.59 (2.96, 4.29)	6.28 (4.77, 8.12)	3.61 (3.04, 4.30)
	Proportion Baseline	6.26 (4.77, 8.05)	3.61 (2.96, 4.33)	6.28 (4.74, 8.02)	3.60 (2.98, 4.34)
	Random Word	6.26 (4.71, 7.76)	3.58 (3.01, 4.31)	6.25 (4.64, 7.96)	3.60 (2.96, 4.33)
	State	6.23 (4.67, 8.01)	3.60 (2.98, 4.32)	6.25 (4.63, 7.94)	3.61 (2.97, 4.34)
XGBoost	All Features	6.22 (4.85, 7.91)	2.18 (1.54, 2.92)	6.24 (4.67, 7.88)	2.16 (1.60, 2.87)
	Baseline	6.13 (4.43, 8.02)	2.91 (2.10, 3.79)	6.11 (4.37, 7.82)	2.92 (2.07, 3.78)
	Characteristics	5.60 (3.97, 7.49)	**2.05 (1.42, 2.77)**	5.61 (4.08, 7.36)	**2.08 (1.44, 2.82)**
	Proportion Baseline	6.98 (5.38, 8.77)	3.44 (2.72, 4.26)	7.02 (5.25, 8.79)	3.47 (2.71, 4.26)
	Random Word	7.16 (5.41, 9.21)	2.82 (1.97, 3.69)	7.18 (5.46, 9.29)	2.84 (1.97, 3.72)
	State	7.30 (5.60, 9.22)	3.33 (2.57, 4.14)	7.31 (5.55, 9.35)	3.33 (2.58, 4.13)

For the **SVM** model, the performance was generally less favorable compared to LightGBM and XGBoost. The best-performing feature set was *All Features*, with an $MAE_{\Delta ADASCOG}$ of 6.29 (95% CI: 4.77, 8.09) and an $MAE_{\Delta MMSE}$ of 3.60 (95% CI: 2.98, 4.29). Despite the inclusion of all feature types, SVM exhibited higher error margins, particularly for MMSE predictions.

For the **XGBoost** model, the combination of *State + Random Word + Characteristics* yielded the most accurate predictions for MMSE, achieving an $MAE_{\Delta MMSE}$ of **2.02** (95% CI: 1.44, 2.76). Similarly, the feature set of *Characteristics* resulted in the best performance for ADAS-Cog, with an $MAE_{\Delta ADASCOG}$ of **5.41** (95% CI: 4.07, 7.11). These results highlight the effectiveness of using state and clinical characteristics in conjunction with embedded information to accurately predict cognitive changes.

As seen in Table 18, the combination of clinical characteristics and specific feature enhancements, such as random words or state embeddings, consistently improved predictive accuracy, particularly for MMSE. However, baseline models and standalone state features generally led to higher error margins, indicating that integrated feature representations are critical for accurate forecasts.

The table presents the performance metrics for predicting annual changes in ADAS-Cog and MMSE (ΔADASCOG and ΔMMSE) across various feature sets, with 95% confidence intervals (CI). In particular, the clinical characteristics feature set used in this analysis excludes current values of the MMSE and ADAS-Cog scores.

For the LightGBM model, the best predictive performance was achieved when using the Characteristics feature set alone, yielding an $MAE_{\Delta ADASCOG}$ of **5.58** (95% CI: 4.11, 7.29) and an $MAE_{\Delta MMSE}$ of **2.55** (95% CI: 1.83, 3.42). These results were the lowest across all LightGBM feature combinations, suggesting that the clinical characteristics features (excluding current MMSE and ADAS-Cog values) effectively capture predictive signals for cognitive decline.

In the XGBoost model, the Baseline + Characteristics feature set demonstrated

Table 19: Performance metrics for predicting changes of ADAS-Cog and MMSE with 95% confidence intervals (CI). The values in parentheses represent the lower and upper bounds of the 95% CI. The metrics represent the predicted annual change (Δ) for MMSE and ADASCOG. Note: This Characteristics feature set does not include the current values of MMSE and ADAS-Cog.

	Metric	**MAE$_{\Delta\text{ADASCOG}}$**	**MAE$_{\Delta\text{MMSE}}$**	**RMSE$_{\Delta\text{ADASCOG}}$**	**RMSE$_{\Delta\text{MMSE}}$**
Model	Feature Set				
LightGBM	All Features	6.21 (4.66, 7.84)	2.61 (1.98, 3.27)	6.20 (4.67, 7.97)	2.63 (2.01, 3.27)
	Baseline	6.76 (4.96, 8.79)	2.58 (1.90, 3.40)	6.70 (5.07, 8.70)	2.56 (1.86, 3.29)
	Characteristics	**5.58 (4.11, 7.29)**	**2.55 (1.83, 3.42)**	**5.65 (4.13, 7.39)**	**2.55 (1.82, 3.38)**
	Proportion Baseline	6.08 (4.64, 7.71)	2.66 (2.01, 3.39)	6.13 (4.55, 7.82)	2.66 (2.01, 3.35)
	Random Word	5.68 (4.28, 7.36)	2.75 (2.13, 3.43)	5.74 (4.26, 7.59)	2.76 (2.13, 3.43)
	State	6.01 (4.49, 7.77)	2.68 (2.04, 3.40)	5.97 (4.37, 7.68)	2.69 (2.05, 3.49)
SVM	All Features	6.32 (4.84, 8.11)	3.61 (3.05, 4.32)	6.25 (4.74, 7.99)	3.63 (3.02, 4.34)
	Baseline	6.22 (4.80, 7.97)	3.61 (3.02, 4.35)	6.25 (4.54, 8.02)	3.62 (2.97, 4.32)
	Characteristics	6.27 (4.72, 8.04)	3.60 (2.95, 4.29)	6.33 (4.69, 8.01)	3.60 (2.98, 4.35)
	Proportion Baseline	6.24 (4.80, 7.93)	3.60 (2.96, 4.33)	6.27 (4.64, 7.87)	3.60 (3.01, 4.33)
	Random Word	6.28 (4.64, 7.96)	3.61 (2.97, 4.33)	6.30 (4.65, 8.00)	3.61 (2.98, 4.35)
	State	6.25 (4.82, 8.02)	3.61 (2.97, 4.32)	6.26 (4.60, 7.92)	3.62 (2.97, 4.34)
XGBoost	All Features	6.77 (5.17, 8.47)	3.36 (2.51, 4.31)	6.78 (5.18, 8.52)	3.34 (2.49, 4.28)
	Baseline	6.15 (4.35, 8.08)	2.91 (2.15, 3.75)	6.10 (4.43, 8.02)	2.92 (2.12, 3.78)
	Characteristics	7.87 (6.05, 9.92)	3.41 (2.53, 4.42)	7.83 (5.99, 9.73)	3.41 (2.44, 4.41)
	Proportion Baseline	6.93 (5.27, 8.74)	3.46 (2.67, 4.36)	6.91 (5.27, 8.80)	3.46 (2.66, 4.32)
	Random Word	7.18 (5.27, 9.17)	2.83 (2.07, 3.66)	7.15 (5.35, 9.19)	2.83 (2.05, 3.72)
	State	7.24 (5.55, 9.10)	3.31 (2.57, 4.07)	7.26 (5.64, 9.11)	3.31 (2.60, 4.08)
LSTM	Location Sequences	27.78 (13.98, 41.57)	27.23 (14.64, 39.82)	35.95 (18.61, 47.31)	34.63 (19.31, 45.00)

competitive performance, with an $\text{MAE}_{\Delta\text{ADASCOG}}$ of 6.36 (95% CI: 4.76, 8.34) and an $\text{MAE}_{\Delta\text{MMSE}}$ of 3.17 (95% CI: 2.31, 4.12). However, using only the Characteristics feature set led to higher errors, with MAE values of 7.87 for ADAS-Cog and 3.41 for MMSE, highlighting the added value of baseline features in this context.

For the LSTM model, which utilized sequential location data, the performance was substantially lower compared to other models. The $\text{MAE}_{\Delta\text{ADASCOG}}$ was 27.78 (95% CI: 13.98, 41.57), and the $\text{MAE}_{\Delta\text{MMSE}}$ was 27.23 (95% CI: 14.64, 39.82). This suggests that while LSTM is adept at modeling sequential data, it may require more extensive feature engineering or larger datasets to achieve competitive performance in predicting cognitive decline.

As summarized in Table 19, excluding current MMSE and ADAS-Cog values has a slight impact on predictive performance; however, specific combinations of characteristics, such as Characteristics alone or Baseline + Characteristics, still yield robust results. These findings highlight the crucial role of carefully selected feature sets in enhancing the accuracy of cognitive decline predictions.

Table 20 summarizes the performance of various models (LightGBM, SVM, XGBoost, and LSTM) in predicting current ADAS-Cog and MMSE scores, evaluated through mean absolute error (MAE) and root mean square error (RMSE), with 95% confidence intervals (CIs) provided in parentheses. The results highlight that feature sets have a significant influence on predictive performance. The LightGBM model achieved its best results using the "Characteristics" feature set, with MAE and RMSE values for ADAS-Cog of 10.60 (95% CI: 8.45, 12.70) and 10.64 (95% CI: 8.61, 12.83), and for MMSE of 4.27 (95% CI: 3.44, 5.14) and 4.28 (95% CI: 3.48, 5.20). Similarly, the XGBoost model performed optimally with the "State + Characteristics" feature set, achieving the lowest MAE for ADAS-Cog of 10.40 (95% CI: 8.12, 12.59) and for MMSE of 4.13 (95% CI: 3.21, 5.15),

Table 20: Performance metrics for predicting current ADAS-Cog and MMSE with 95% confidence intervals (CI). The values in parentheses represent the lower and upper bounds of the 95% CI.

Model	Metric / Feature Set	$\mathbf{MAE_{ADASCOG}}$	$\mathbf{MAE_{MMSE}}$	$\mathbf{RMSE_{ADASCOG}}$	$\mathbf{RMSE_{MMSE}}$
LightGBM	All Features	12.96 (10.39, 15.48)	4.90 (3.98, 5.82)	12.99 (10.74, 15.49)	4.92 (4.04, 5.78)
	Baseline	12.38 (10.33, 14.65)	4.73 (3.92, 5.61)	12.27 (10.05, 14.44)	4.75 (3.94, 5.58)
	Characteristics	**10.60 (8.45, 12.70)**	4.27 (3.44, 5.14)	10.64 (8.61, 12.83)	4.28 (3.48, 5.20)
	Proportion Baseline	11.57 (9.37, 13.82)	4.79 (3.96, 5.73)	11.60 (9.40, 13.97)	4.80 (3.99, 5.71)
	Random Word	10.77 (9.01, 12.67)	4.40 (3.61, 5.23)	10.76 (8.92, 12.53)	4.35 (3.59, 5.17)
	State	10.64 (8.67, 12.70)	**4.24 (3.40, 5.14)**	**10.60 (8.60, 12.69)**	**4.24 (3.37, 5.10)**
SVM	All Features	10.85 (8.41, 13.28)	4.71 (3.79, 5.71)	10.77 (8.37, 13.22)	4.71 (3.84, 5.62)
	Baseline	10.76 (8.38, 13.28)	4.51 (3.60, 5.45)	10.66 (8.10, 13.13)	4.48 (3.62, 5.36)
	Characteristics	10.89 (8.59, 13.56)	4.44 (3.46, 5.42)	10.90 (8.72, 13.32)	4.46 (3.53, 5.45)
	Proportion Baseline	10.72 (8.47, 13.15)	4.46 (3.63, 5.44)	10.69 (8.29, 13.34)	4.47 (3.61, 5.41)
	Random Word	10.77 (8.52, 13.25)	4.40 (3.49, 5.41)	10.79 (8.51, 13.30)	4.39 (3.52, 5.32)
	State	10.62 (8.19, 13.08)	4.42 (3.50, 5.38)	10.72 (8.26, 13.12)	4.39 (3.49, 5.33)
XGBoost	All Features	11.84 (9.50, 14.12)	4.86 (3.95, 5.82)	11.82 (9.40, 14.32)	4.89 (3.96, 5.81)
	Baseline	11.64 (9.00, 14.39)	5.66 (4.51, 6.75)	11.60 (8.89, 14.58)	5.65 (4.62, 6.71)
	Characteristics	11.77 (8.70, 14.82)	5.03 (3.76, 6.28)	11.74 (9.07, 14.62)	5.02 (3.74, 6.38)
	Proportion Baseline	11.57 (9.22, 14.09)	5.29 (4.26, 6.33)	11.63 (8.99, 14.26)	5.28 (4.31, 6.32)
	Random Word	12.61 (9.81, 15.62)	4.79 (3.77, 6.00)	12.66 (9.97, 15.68)	4.81 (3.78, 5.96)
	State	11.09 (8.53, 13.94)	4.67 (3.56, 5.89)	10.95 (8.38, 13.57)	4.65 (3.62, 5.69)
LSTM	Location Sequences	11.35 (8.56, 14.13)	4.93 (3.94, 5.91)	14.51 (11.66, 16.89)	5.99 (4.76, 7.02)

along with the lowest RMSE for ADAS-Cog of 10.44 (95% CI: 8.38, 12.65) and for MMSE of 4.13 (95% CI: 3.19, 5.07). The SVM model showed competitive performance with the "State" feature set, achieving MAE values of 10.62 (95% CI: 8.19, 13.08) for ADAS-Cog and 4.42 (95% CI: 3.50, 5.38) for MMSE, and RMSE values of 10.72 (95% CI: 8.26, 13.12) for ADAS-Cog and 4.39 (95% CI: 3.49, 5.33) for MMSE. In contrast, the LSTM model, which used location sequences as input, demonstrated relatively higher RMSE values of 14.51 (95% CI: 11.66, 16.89) for ADAS-Cog and 5.99 (95% CI: 4.76, 7.02) for MMSE, suggesting the need for further optimization. Overall, models utilizing feature sets enriched with "Characteristics" and "State + Characteristics" consistently outperformed those relying on baseline or random word features. The inclusion of confidence intervals highlights the robustness of these results in predicting cognitive metrics.

A.8.1 Ablation Study: Impact of Time-Series Length on Model Performance

To investigate how the length of the time-series input affects the model's predictive performance, we conducted an ablation study by varying the time durations provided to the ridge model. This analysis focuses specifically on assessing the impact of input data collected over different periods: 7 days, 15 days, 30 days, 90 days, and 180 days. By analyzing these variations, we aim to understand the optimal time frame and seasonal changes for aggregating data from daily activities to predict cognitive scores effectively.

As shown in Figure 40, which visualizes the prediction errors over varying time durations, we observed that the model generally achieves lower prediction errors when using longer time series inputs (30 to 180 days). Specifically, for both MAE and RMSE metrics, the model's performance slightly degraded as the time duration was less than 30 days. This trend was consistent across all metrics, suggesting that more participant movement data has a greater predictive value for cognitive change.

When excluding the current ADAS-Cog and MMSE values from the set of characteristics, as shown in Figure 39, the error values showed a similar pattern. The absence of

current scores slightly increased MAE and RMSE, reinforcing the importance of longitudinal behavioral data for accurate predictions.

A.8.2 Prediction of Absolute Cognitive Scores (ADASCOG and MMSE)

In contrast, Figure 41, which represents the performance metrics for predicting absolute ADAS-Cog and MMSE scores. When predicting the current value of cognitive scores, MAE and RMSE remained relatively stable over time durations, with only minor decreases in error as the time frame extended to 180 days. This stability suggests that longer time series data can be valuable in providing information on cognitive status.

A.8.3 State Feature Performance

For this ablation study, we observed that as the input time span increases, the performance gap between the state model and the baseline model widens. This finding further supports the notion that the state feature, derived from our in-depth analysis of patient mobility patterns, captures significant information on patient behavioral patterns and cognitive abilities. Furthermore, with longer input time series, the state model demonstrates greater robustness in predicting cognitive function compared to the baseline, highlighting its ability to capture long-term temporal information, particularly when the time series is exceptionally long.

Moreover, in nearly all experiments, the performance of the state feature, which only depicts the movement of patients, is comparable to that of the characteristic feature. In the case of extended time series, the state feature even outperforms the characteristic feature in terms of prediction accuracy. This indicates that the state feature we generate, which could provide a deeper insight into clinical diagnostic results, may be an effective predictor of cognitive status.

A.9 Detailed Data Extracting and Synthetic Data Generation Workflow

Figure 11 illustrates the workflow for processing real-world time series data and generating synthetic data using Retrieval-Augmented Generation (RAG). The pipeline is divided into several stages, which are described below.

A.9.1 Raw Data Representation

The raw data consists of participant movement trajectories across predefined locations (e.g. — Wake —, — Sleep —, — Lounge —, — Bedroom —, — Bathroom —, — Hallway —, — Kitchen —, — Nowhere —). Each data point represents the location of the participant over a 24-hour period, with a temporal resolution of 20 minutes. The raw data contains irregularities and gaps that are addressed in subsequent steps.

A.9.2 Rectification of Time-Series Data

To address inconsistencies in the raw data, a rectification process is applied.

- The raw time-series data is smoothed and interpolated to fill missing values.

- Temporal gaps are corrected by aggregating and aligning data points in regular 20-minute intervals.

This results in a corrected participant location time series, preserving key behavioral patterns while ensuring data consistency.

A.9.3 Conversion to Structured Text

The rectified time-series data is converted into structured text for compatibility with the RAG framework. Each sequence of location data is represented as a string of location names separated by commas. This textual format enables seamless integration with pre-trained language models and vector databases.

A.9.4 Retrieval-Augmented Generation (RAG) Query and Generation

The structured text is used as input for the RAG framework. This stage involves:

- **Querying the Vector Database:** The input text is embedded using a BERT-based pre-trained model, fine-tuned on clinical movement data. These embeddings are queried against a vector database to retrieve relevant movement patterns.
- **Conditioned Generation:** The retrieved patterns, along with additional clinical information from the patient (e.g., UTI status, MMSE scores, ADAS-COG, HADS), are fed into a language model to generate synthetic movement trajectories. The generation process incorporates both the retrieved embeddings and the condition prompt to ensure clinical plausibility.

A.9.5 Synthetic Data Rectification

The generated structured text is converted back to a time series format, mirroring the rectification process applied to the original data. This results in a rectified participant location time series for the synthetic data, which closely resembles the real data in terms of temporal fidelity and location transitions.

A.10 Base Generation Model Architecture and Parameters

For the dataset generation, we employed the `Llama-3.2-3B-Instruct` model, a large-scale instruction-tuned language model optimized for conditional text generation. The model was fine-tuned on the extracted embeddings and trained to generate realistic movement sequences based on historical trajectory patterns.

The text generation pipeline was configured with the following hyperparameters:

- **Precision:** bfloat16
- **Temperature:** 0.9 (moderate diversity in generation)
- **Top-k Sampling:** 100 (limits token choices to the top 100 with highest probability)
- **Top-p Sampling:** 0.5 (nucleus sampling for controlled randomness)
- **Maximum Tokens:** 512 (limits the length of generated sequences)

- **Minimum Length:** 50 (ensures sufficient content per generation)
- **Early Stopping:** Enabled (stops generation once a complete sentence is formed)
- **Sampling Method:** Enabled (introduces variation instead of deterministic greedy search)

The model was deployed on an A100 80GB GPU, running for 390 hours, 9 minutes, and 44 seconds to generate the complete dataset. This extensive computation ensured that the generated trajectories adhered to realistic behavioral patterns derived from the training data while preserving the structural consistency of the original dataset.

A.11 Downstream Task Model Data Preparation

This section describes the construction and parameters of the models used in the experiments, including an LSTM-based model and a Transformer-based model. In addition, the cross-validation strategy and performance evaluation methods are detailed.

A.11.1 Data Preparation

The input data for the models consisted of clinical movement records from the last 7 or 30 days, depending on the specific task. These records were converted into continuous movement vectors by concatenating sequential embeddings over the selected time window. Clinical experts guided the choice of time window (7 or 30 days) to ensure the data's relevance for each downstream task. For the agitation task, we use the patient's movement trajectory from the seven days preceding the recorded event as input. In contrast, for UTI, ADAS-Cog, MMSE, and NPI, we used the patient's movement trajectory from the thirty days preceding the recorded event as the input time series. For classification tasks, we applied random enhancement to the minority class to mitigate the class imbalance.

A.11.2 Task Participant Profiling

For MMSE, the entire dataset contains 123 valid records from 87 participants. For NPI, there are 318 valid records from 89 participants. For ADAS-Cog, the data set includes 260 valid records from 94 participants. After applying data augmentation, the dataset contains 602 valid records from 67 participants for agitation and 656 valid records from 82 participants for UTI.

A.11.3 Model Architecture

Please note that the architectures, parameters, and training methodologies for all regression and classification models are as follows. In our study, we implemented both regression and classification models with similar base architectures but with modifications tailored to each task. For the regression tasks, two distinct model architectures were employed: an LSTM-based model and a Transformer-based model. The LSTM model comprises an LSTM layer with 64 hidden units (with an input shape of 31 by 576), followed by a dropout layer with a dropout rate of 0.1 to mitigate overfitting, a dense layer with 20 neurons using ReLU activation, and a final dense layer with linear activation to output continuous predictions. The Transformer model follows a hybrid

architecture that integrates a Transformer encoder with a fully connected multilayer perceptron (MLP) for final prediction. The input features, represented as encoded vectors, are first processed through a dense layer with 128-dimensional embeddings, incorporating regularization L2 (L2 penalty of 1×10^{-4}). A Transformer encoder follows, consisting of four self-attention heads and an intermediate feedforward network with 128 hidden units, utilizing a dropout rate of 0.1 to prevent overfitting. The encoded representations are aggregated using global average pooling to obtain a fixed-length feature vector. To ensure compatibility between the Transformer output and subsequent MLP layers, batch normalization is applied. The MLP component follows a hierarchical structure with progressively decreasing hidden dimensions (64, 32, 20, and 20 neurons, respectively), each employing ReLU activation and regularization of L2 to encourage sparsity and stable representations. In addition, a dropout layer (0.1 probability) is introduced to mitigate overfitting. The output layer consists of a single neuron with a linear activation function, making it suitable for regression tasks. Both models are compiled using the Adam optimizer and use mean absolute error as the loss function; training is conducted using leave one out(LOO) strategy with early stopping based on validation loss, and performance is evaluated by computing mean absolute error (MAE) and mean squared error (MSE) along with their 95% confidence intervals. For the classification tasks, similar architectural frameworks were adopted, with modifications to accommodate binary outcomes. Specifically, the final dense layer is replaced with one that uses sigmoid activation, and the loss function is changed to binary cross-entropy, while incorporating accuracy as an evaluation metric. Data preparation for classification includes converting patient data into binary labels (e.g., where a label of 1 indicates the occurrence of agitation).

A.11.4 Calibration Results on Regression Tasks

Figure 42, 43, and 44 compare the predictions (prognosis) and ground truth values on the test set for three downstream regression tasks using real and generated datasets. It can be observed that some high values are not fully fitted, but in most cases, the predictions fall within a reasonable range. It demonstrates that the model's predictions do not suffer from severe overfitting or underfitting, such as always predicting the same value.

A.12 Generation Model Ablation Study

In this section, our objective is to evaluate the influence of the chosen generative model on the experimental outcomes. Specifically, we selected two additional models: **Qwen2.5-3B-Instruct** and **Gemma-2B-IT**. These models were chosen because their model sizes are comparable and compatible with our available computational resources. Importantly, the generation process for all models employed identical prompts and hyperparameter settings, ensuring a fair comparison. For Qwen2.5-3B-Instruct, generating all 66,096 records took approximately 287.2 hours, and for Gemma-2B-IT, generating all 66,096 records took approximately 537.55 hours.

The results, as shown in Figure 45, indicate that the choice of generative model had only a minor impact on the results, with no substantial changes observed in key performance trends. This suggests that the performance of the downstream task is relatively robust to the selection of the specific generative model, as long as the generation process is controlled under consistent conditions.

A.13 Visualization Design for Clinical Trajectory Chat Framework

Figure 46 integrates user inputs, system prompts, RAG (retrieval-augmented generation) modules, and model-generated outputs within an intuitive layout.

A.13.1 1. Dialogue Components

The visualization consists of the following key elements:

- **System Message:** The system displays a condition prompt that describes the input data, patient profile, and the specific task. For example, input data include time series movement trajectories across predefined locations (e.g., —Bedroom—, —Kitchen—, —Lounge—), and the task is to generate privacy-preserving synthetic trajectories while maintaining clinical plausibility.
- **RAG Module:** Embeddings for clinical movement trajectories are generated using a BERT-based model fine-tuned on clinical datasets. These embeddings are queried against a vector database to retrieve the 30-day trajectories of this record, which are displayed as part of the visualization.
- **User and Model Interaction:** User messages (e.g., "Generate synthetic sequences") are displayed on the left, while the corresponding model responses are displayed on the right. The model output integrates retrieved trajectories and clinical information to generate plausible synthetic trajectories, annotated with key features such as patient conditions (e.g., vascular dementia, a history of UTI).
- **Annotations:** Annotation boxes highlight specific transitions or behaviors within the generated trajectory, such as potential wandering patterns or significant movement transitions.

A.13.2 2. Example Usage

An example of the framework usage is as follows:

- **Input:** A condition prompt specifies a patient profile (e.g. 84 years, vascular dementia, UTI positive) and requests synthetic trajectory generation.
- **RAG Retrieval:** The top-3 trajectories retrieved from the vector database include transitions such as —Lounge— $\rightarrow$ —Hallway— $\rightarrow$ —Bedroom—.
- **Output:** The model generates a plausible synthetic trajectory while preserving privacy and clinical plausibility. Key patterns (e.g., wandering behavior) are annotated and highlighted for further analysis.

This visualization framework offers a structured and intuitive approach to presenting the interaction among users, retrieval-augmented generation, and model output. It is particularly useful for demonstrating synthetic data generation workflows in academic and clinical settings.

A.13.3 Clinical Challenges

From a clinical perspective, aligning the outputs of AI models with actionable insights for healthcare providers is a key challenge. The proposed low-rank state representations must closely correlate with clinically significant metrics, such as MMSE and ADAS-Cog scores, to support diagnosis and intervention planning. However, the limited size and scope of our dataset, which includes only 134 participants with complete data from 50 patients, presents a significant limitation. This small sample size may not fully capture the diversity of movement and behavioral patterns observed in the larger population of people living with dementia. As a result, the generalizability of the model to other settings or populations may be compromised. Furthermore, variability in patient behavior due to external factors, such as seasonal changes or caregiver interactions, requires models that can differentiate between pathological changes and normal variations. Integrating AI-driven insights into existing clinical workflows without increasing the burden on healthcare providers remains a critical consideration to ensure the practical utility of these systems.

Addressing these challenges requires an interdisciplinary approach, incorporating the expertise of AI researchers, clinicians, ethicists, and regulatory bodies, to develop robust, equitable, and clinically impactful healthcare monitoring solutions. Expanding the data set to include a more diverse and representative sample of patients will also be essential to improve the reliability and applicability of the model.

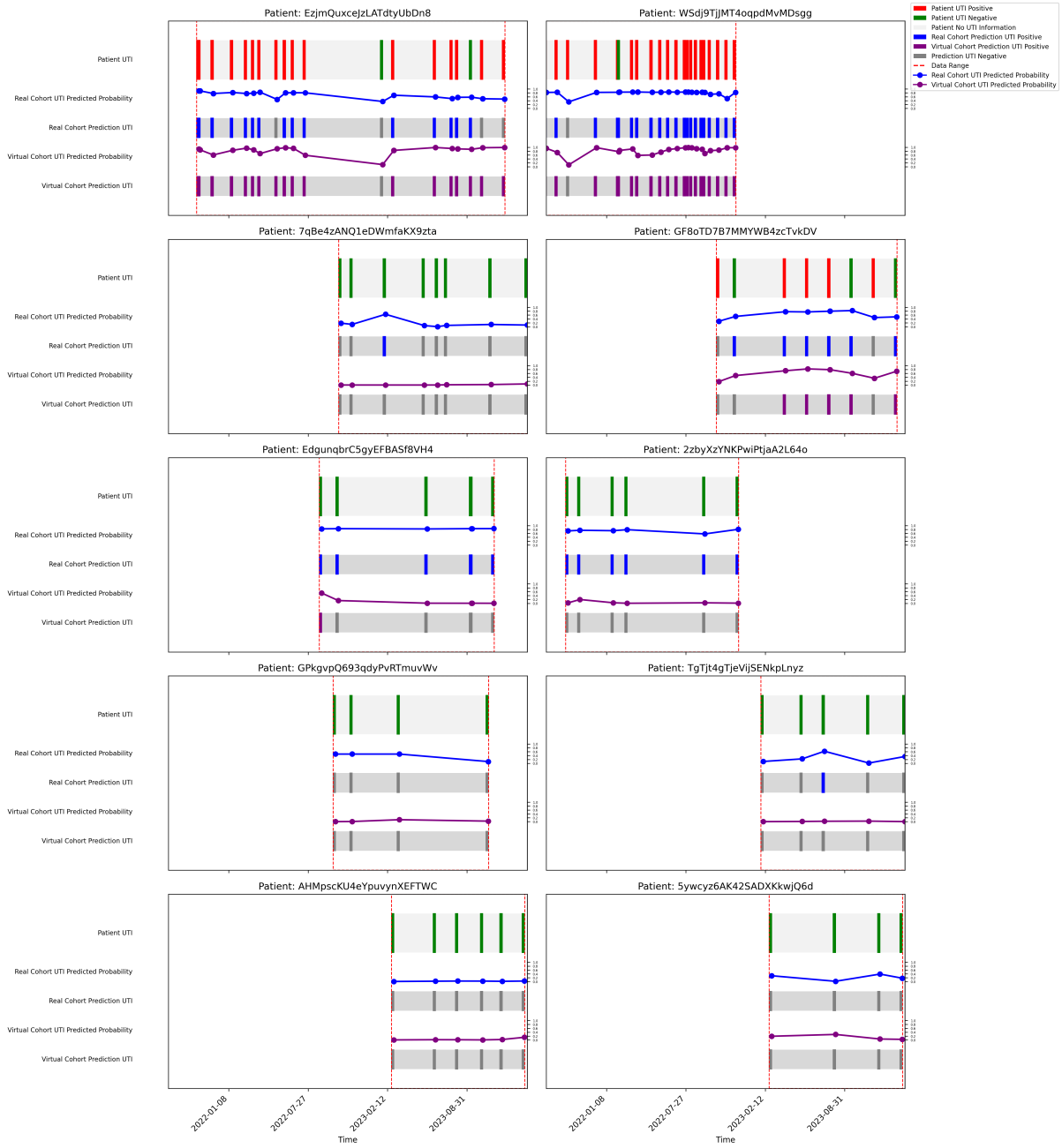


Figure 25: UTI classification results on the test set, comparing real and virtual cohort predictions across multiple patients.

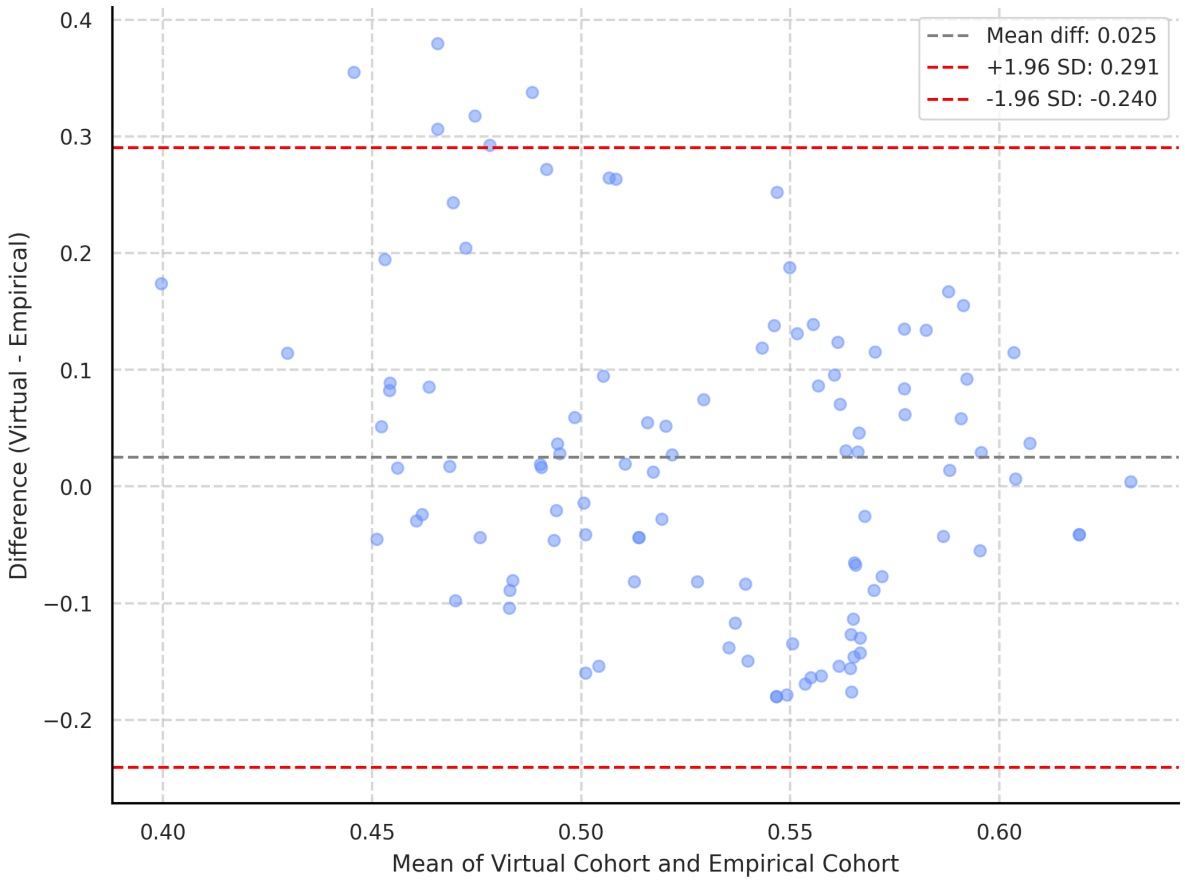


Figure 26: Bland–Altman for Agitation

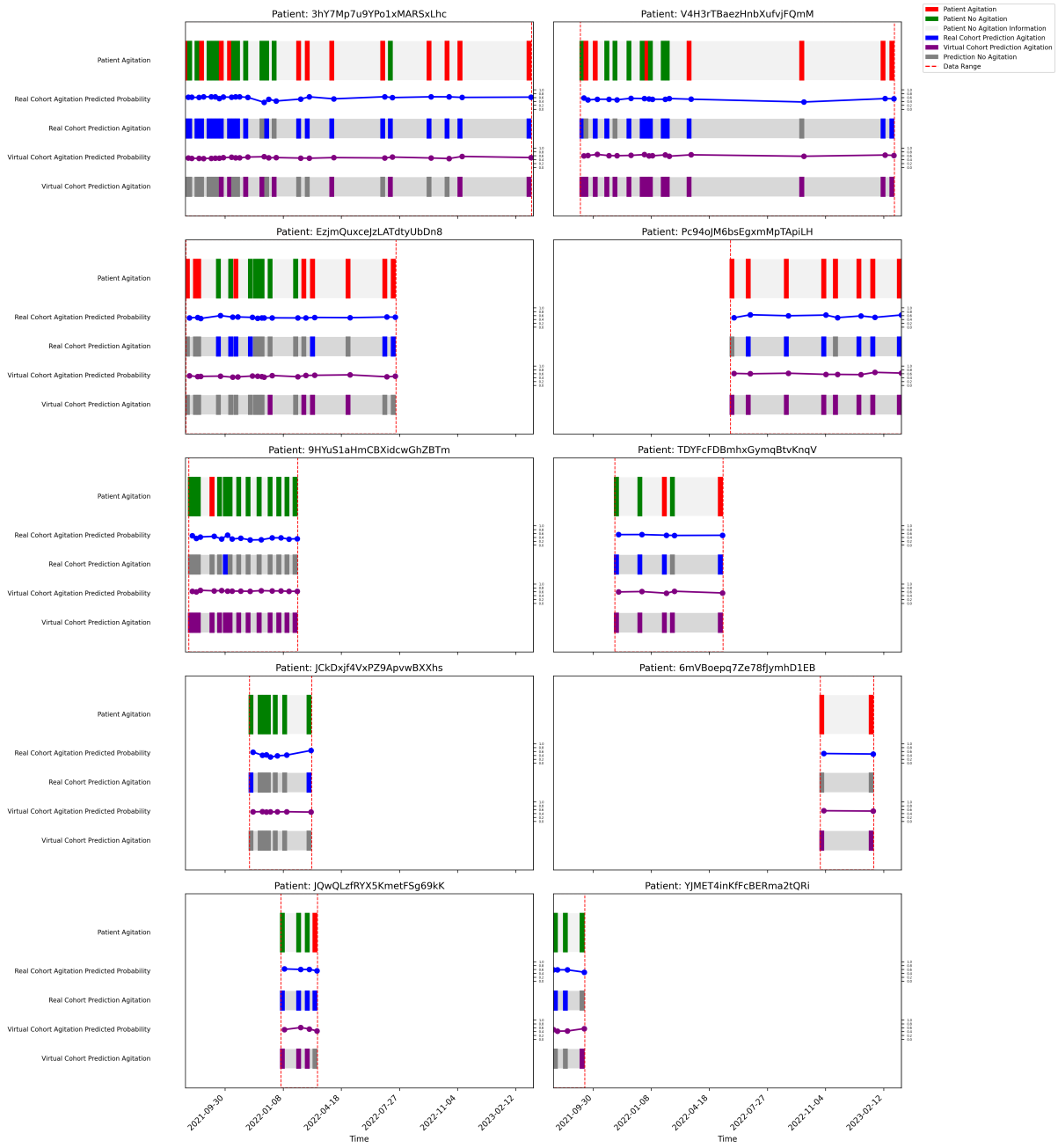


Figure 27: Agitation classification results on the test set, comparing real and virtual cohort predictions across multiple patients.

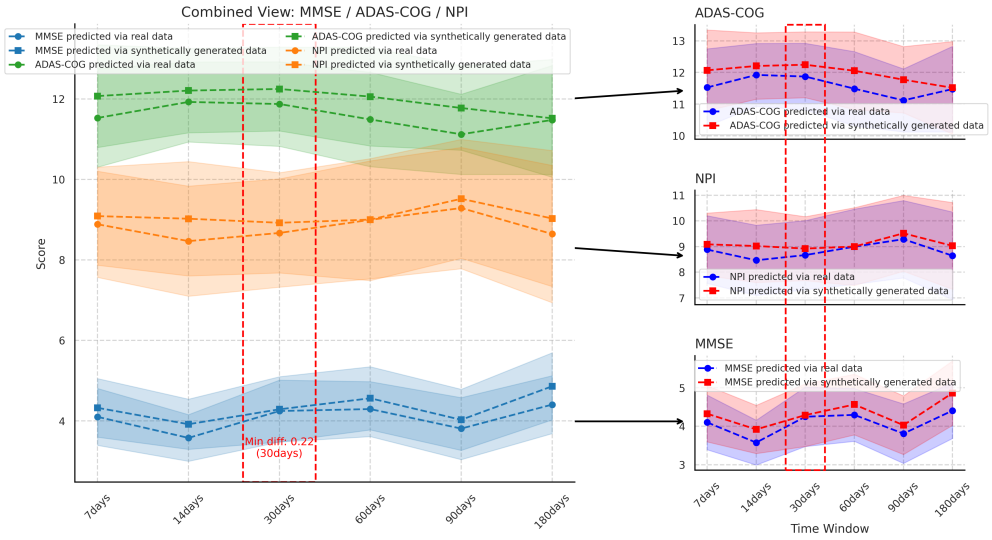


Figure 28: MAE Comparison of Real vs. Virtual Cohorts(Largest & Smallest Diff)

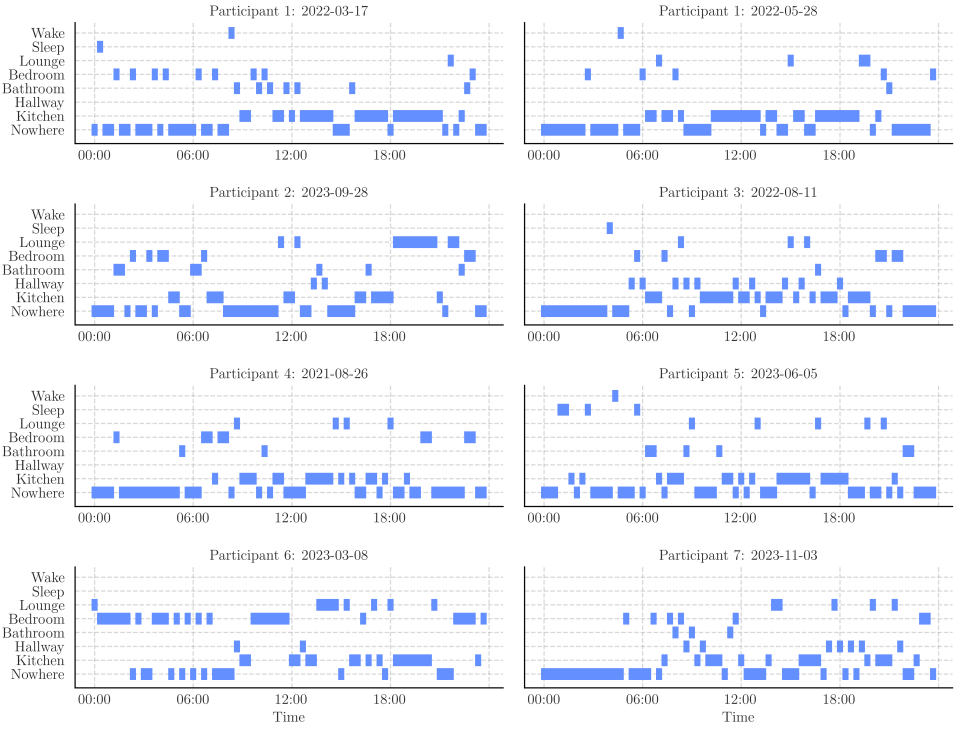


Figure 29: Location timeseries for articipants in cluster 1



Figure 30: Location timeseries for participants in cluster 2



Figure 31: Location timeseries for participants in cluster 3



Figure 32: Location timeseries for participants in cluster 4

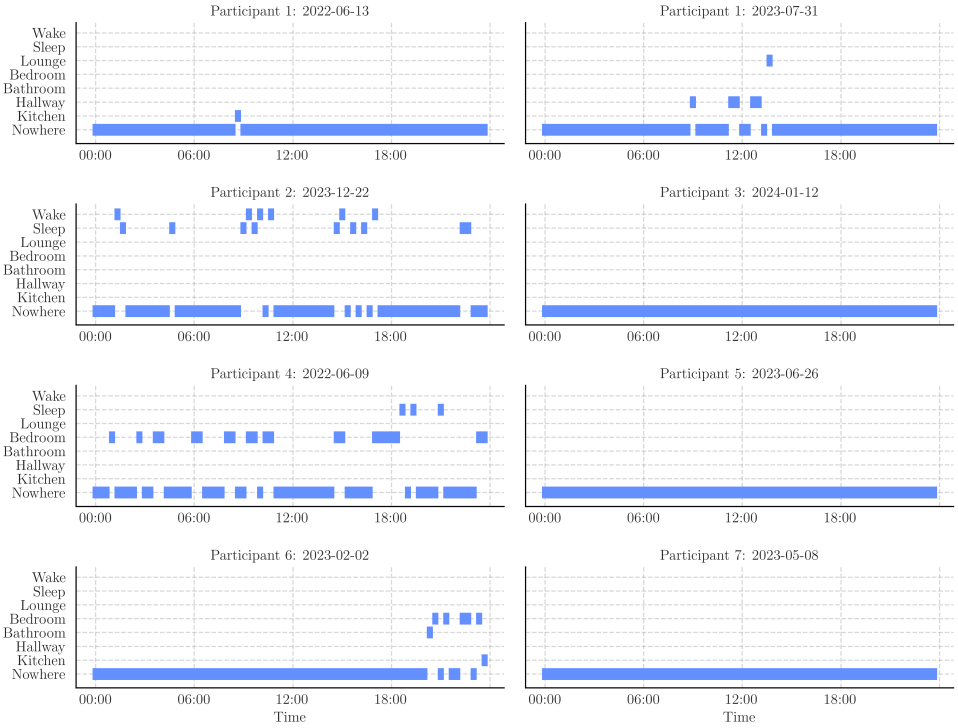


Figure 33: Location timeseries for participants in cluster 5

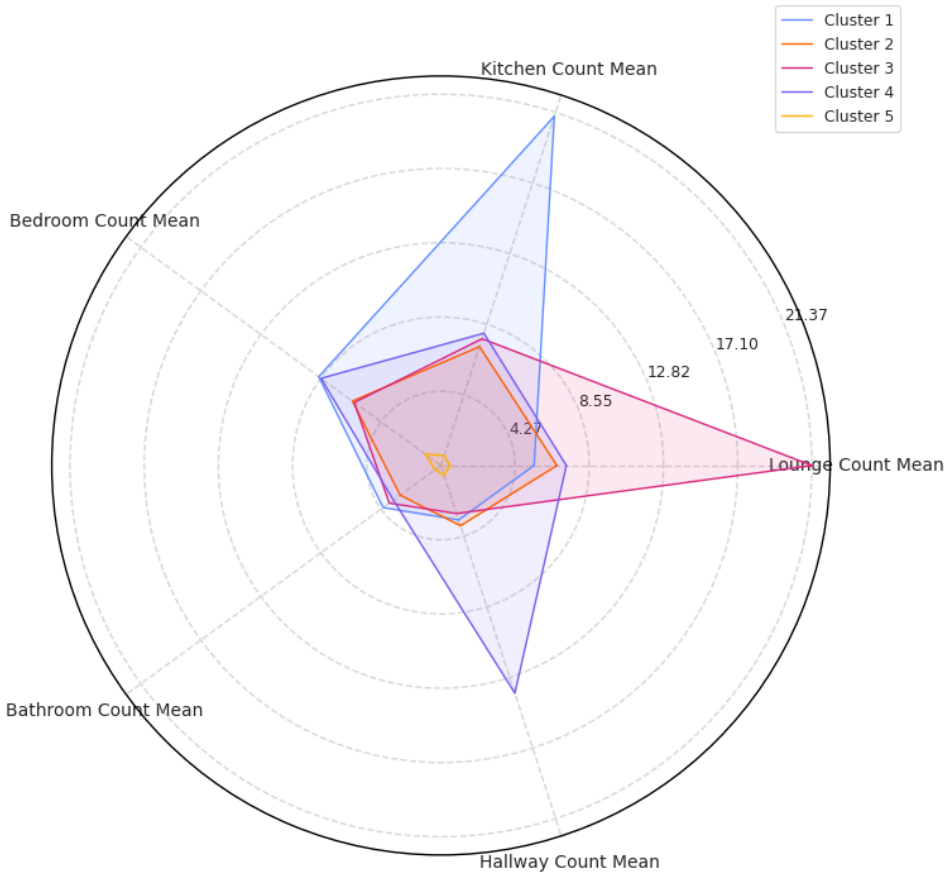


Figure 34: Mean of location count in five clusters

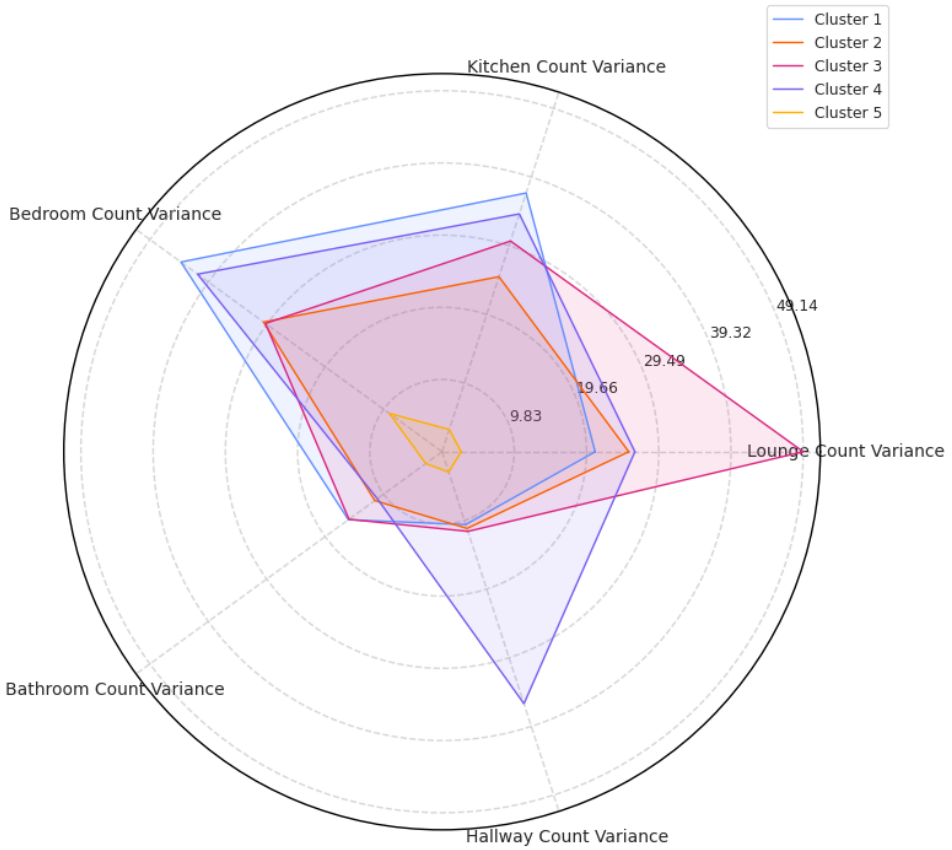


Figure 35: Variance of location count in five clusters

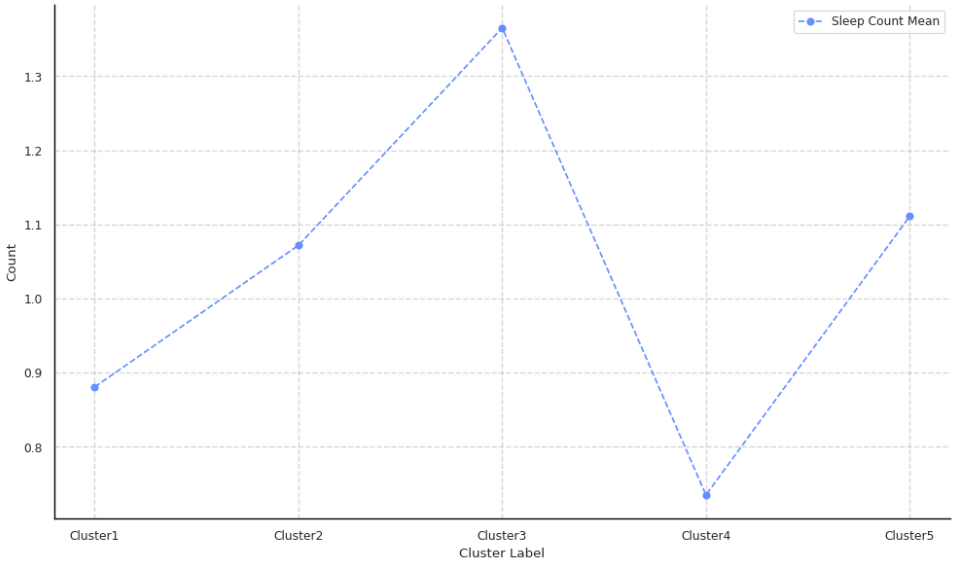


Figure 36: Mean of wake count in five clusters

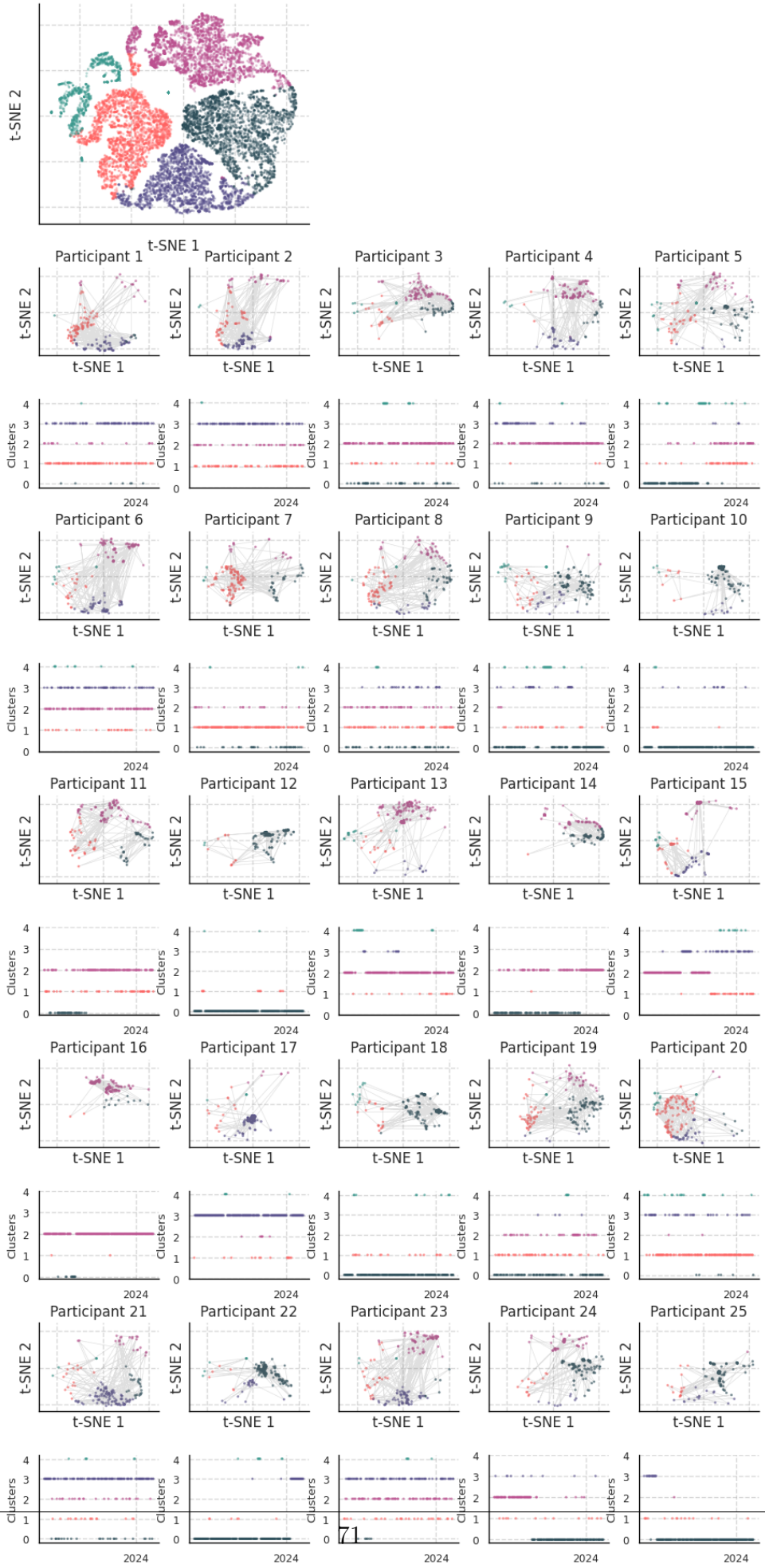


Figure 37: T-SNE for individuals in test set

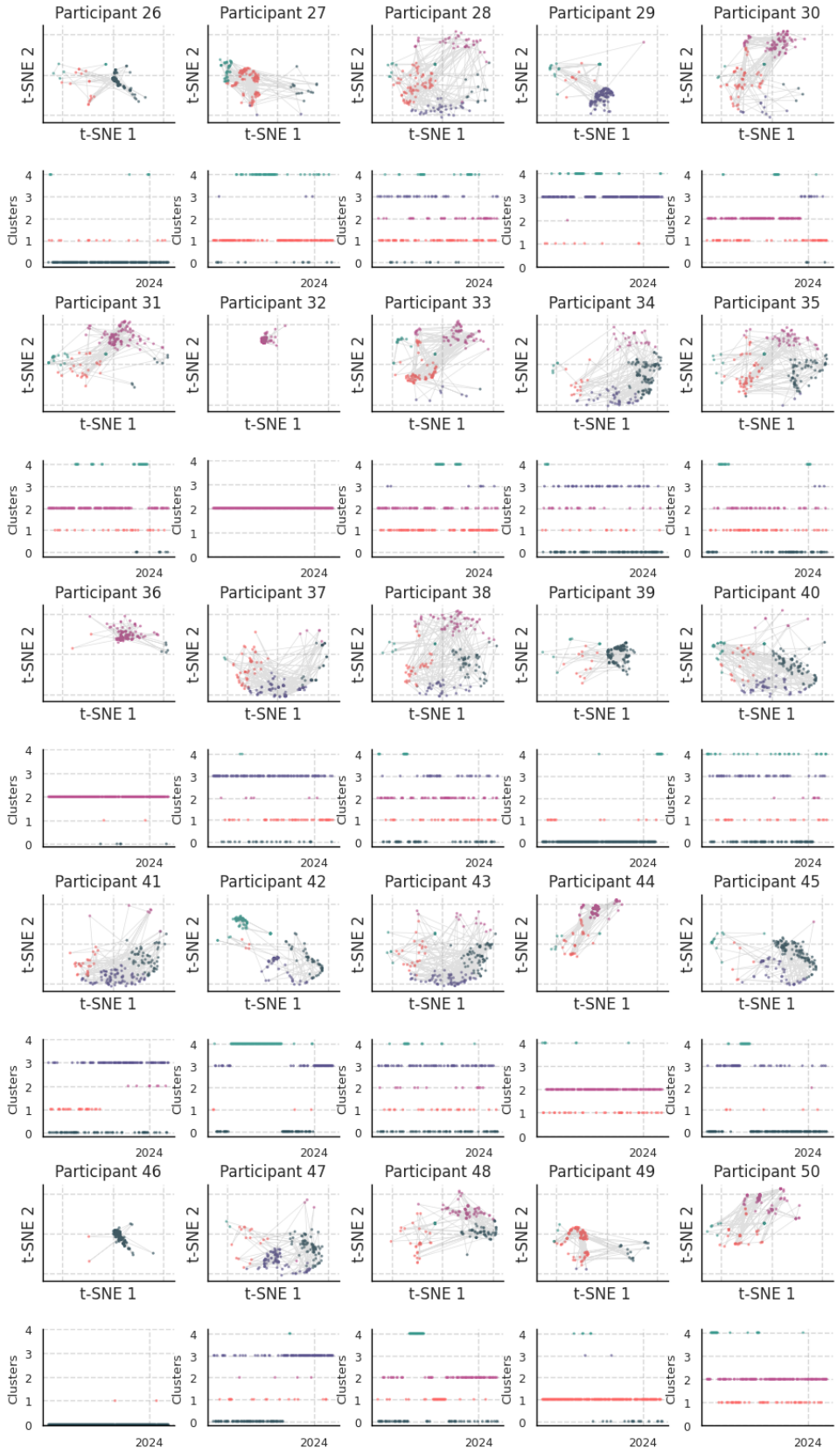


Figure 38: T-SNE for individuals in test set

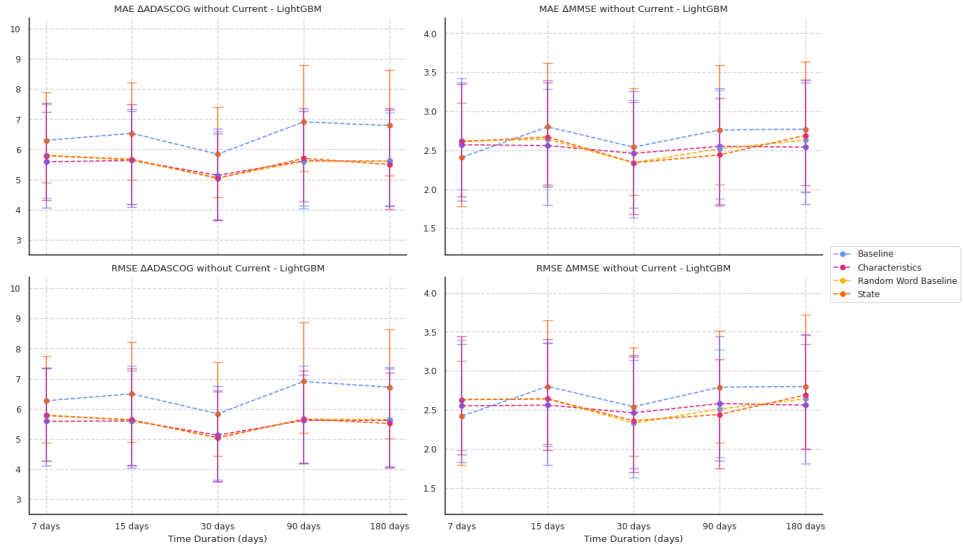


Figure 39: Prediction error for cognitive change (ΔADAS-Cog and ΔMMSE) without current cognitive score in feature set. This figure illustrates the impact on prediction errors (MAE and RMSE) for cognitive change when excluding the current cognitive score from the feature set across different time durations.

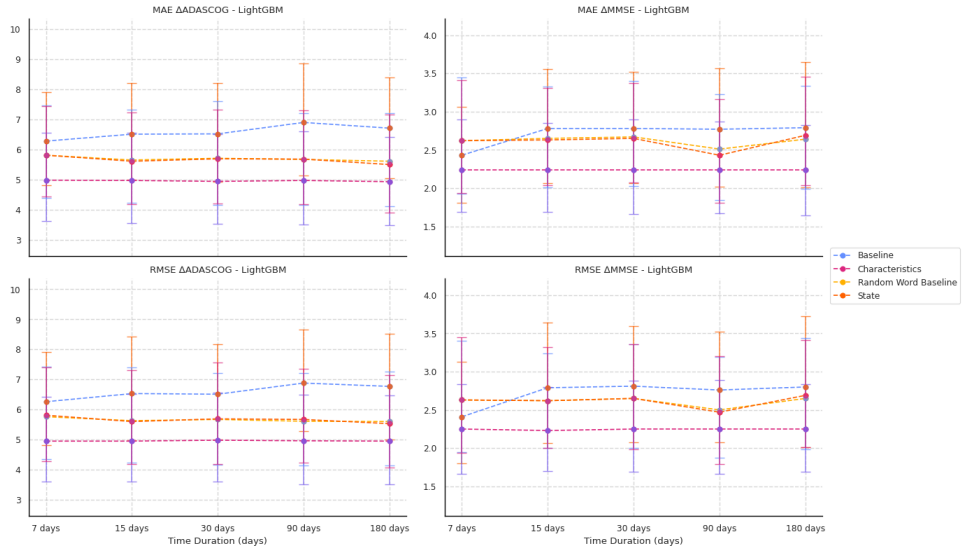


Figure 40: Prediction error for cognitive change (ΔADAS-Cog and ΔMMSE). This figure shows the MAE and RMSE for cognitive change predictions using varying time-series lengths, including baseline, characteristics, and state feature sets.

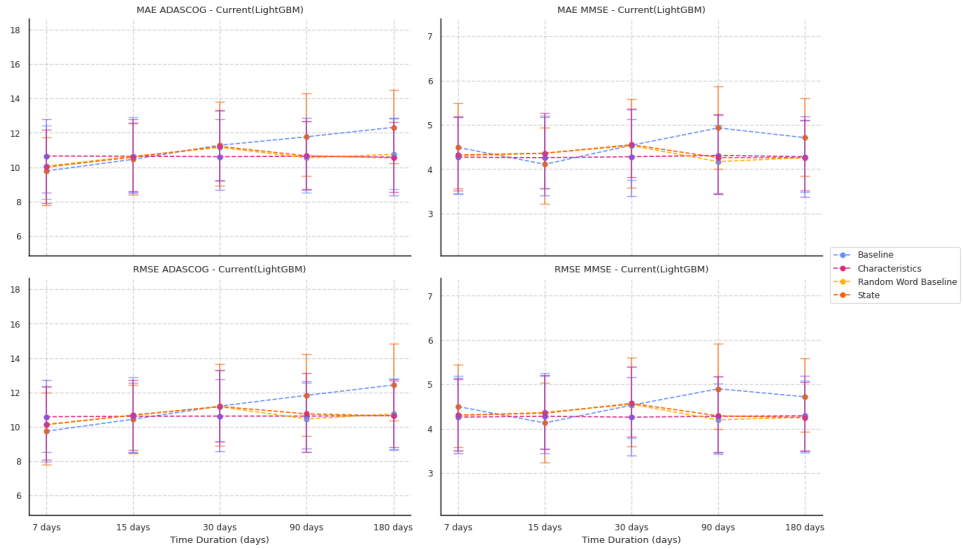
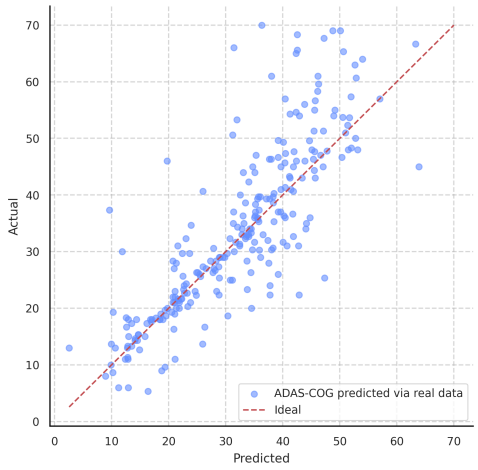
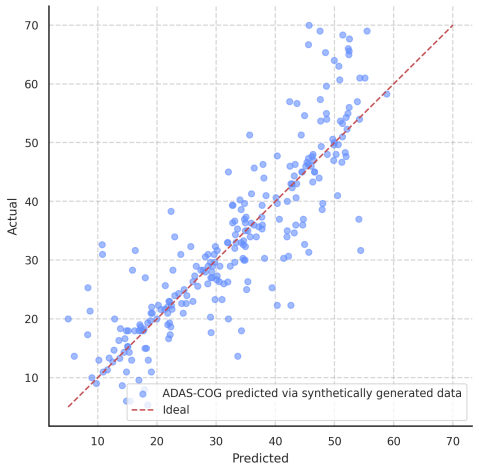


Figure 41: Prediction error for current cognitive scores (ADAS-Cog and MMSE). This figure displays the MAE and RMSE for absolute cognitive score predictions, illustrating the performance stability across various time-series lengths with baseline, characteristic, and state feature sets.

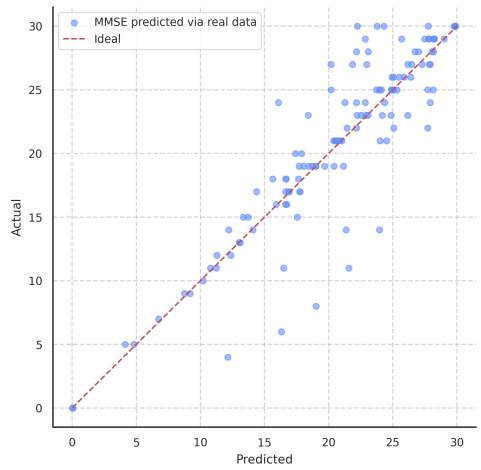


(a) Calibration plot for real dataset (ADAS-COG)

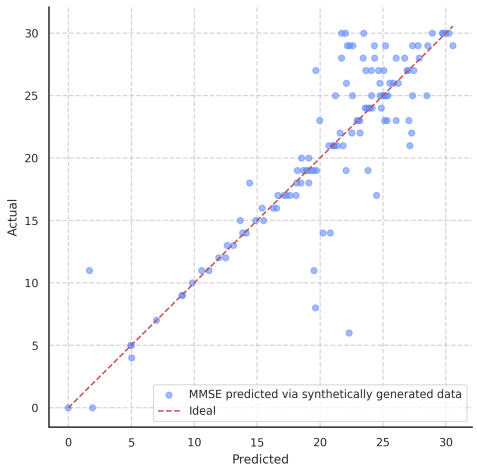


(b) Calibration plot for synthetic dataset (ADAS-COG)

Figure 42: Comparison of calibration plots for ADAS-COG task

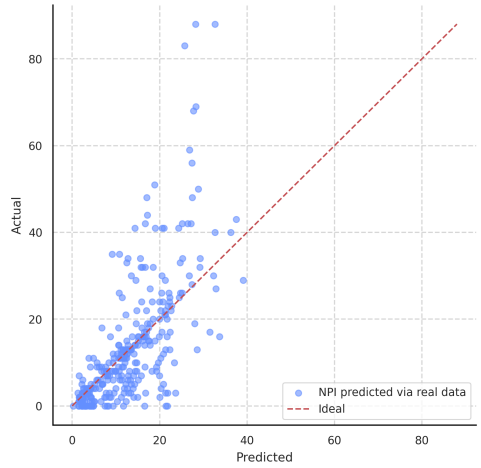


(a) Calibration plot for real dataset (MMSE)

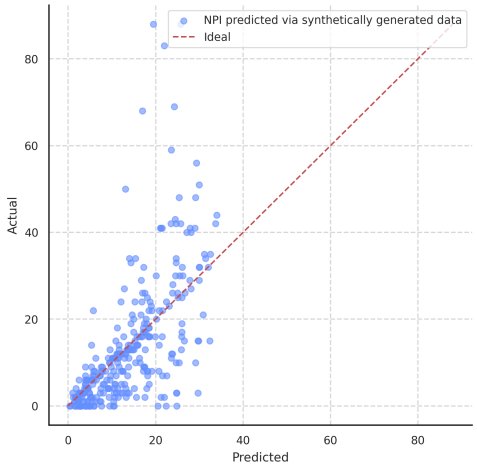


(b) Calibration plot for synthetic dataset (MMSE)

Figure 43: Comparison of calibration plots for MMSE task



(a) Calibration plot for real dataset (NPI)



(b) Calibration plot for synthetic dataset (NPI)

Figure 44: Comparison of calibration plots for NPI task

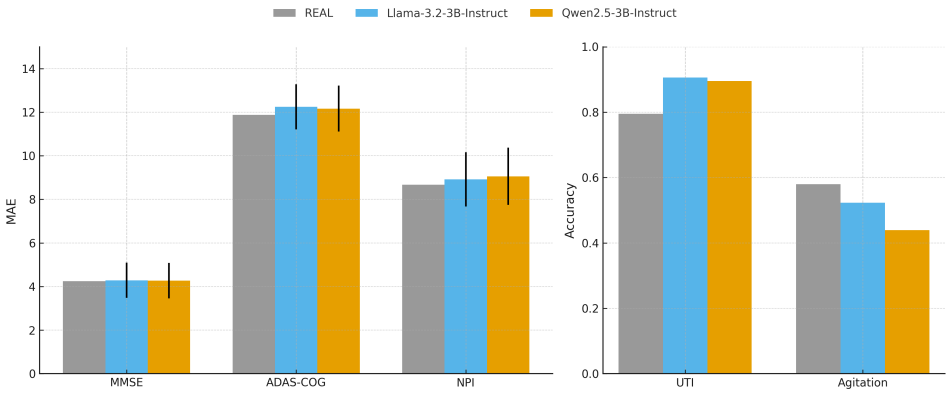


Figure 45: Performance comparison across different generative models. The results show that the selection of the generative model has limited impact on key performance trends.

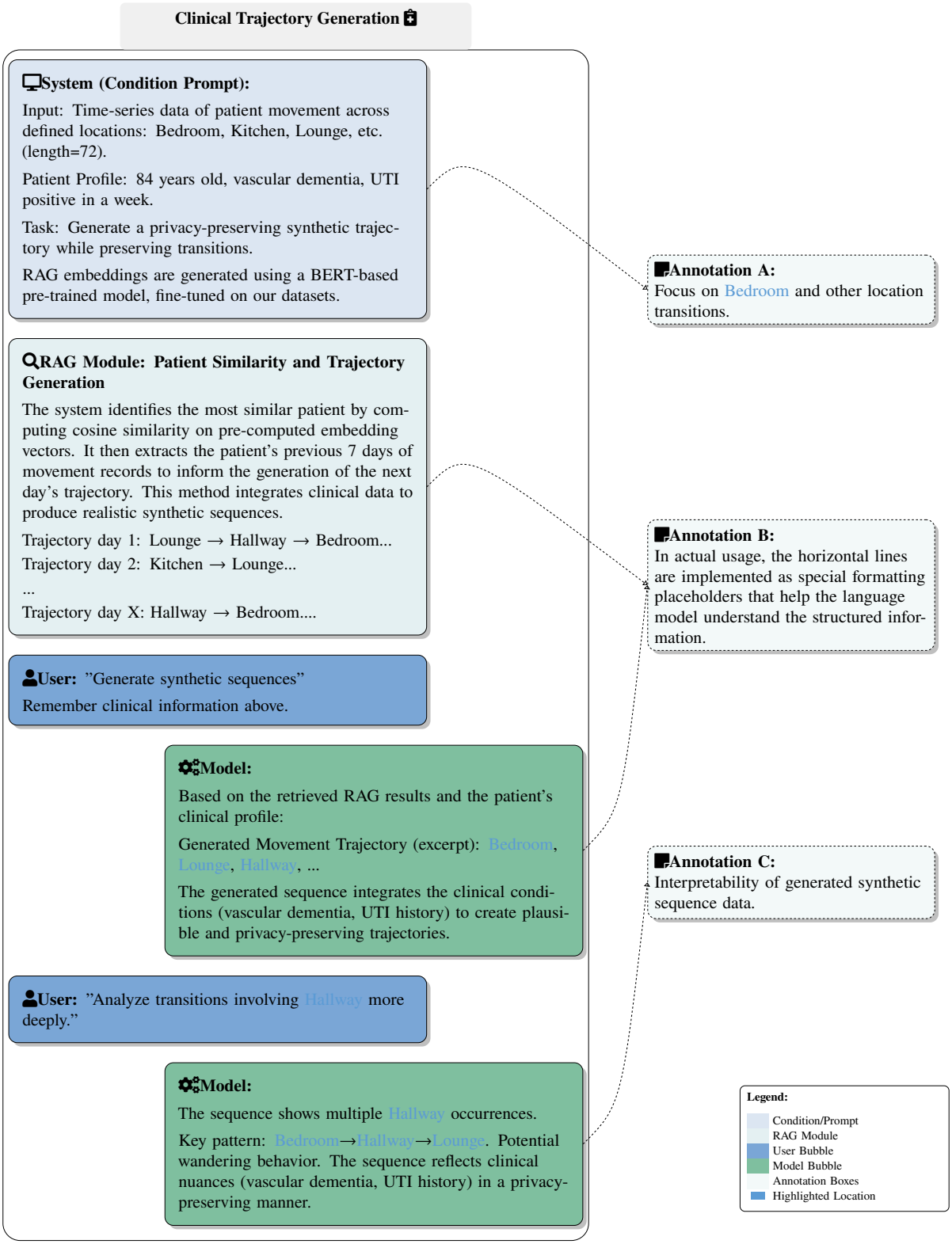


Figure 46: Clinical Sequence generator example (LLM)