**Report on Synthetic Data Generation-BHI 24 Track 1**

**Introduction**

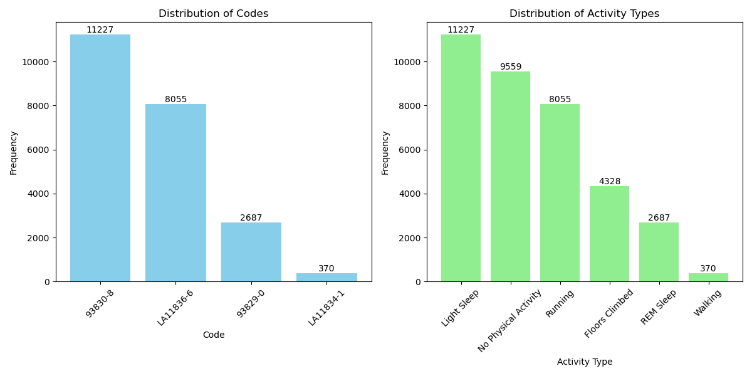
This report outlines the process of generating and evaluating synthetic new patient activity data from 100 patients, focusing on various types of activities, including light sleep, REM sleep, running, walking, no physical activity, and floors climbed. The goal is to create a high-quality synthetic dataset that mimics the statistical distribution of real patient data, enabling further research and analysis in activity recognition and health monitoring.

**Method**

**Step 1: Statistical Distribution Analysis**

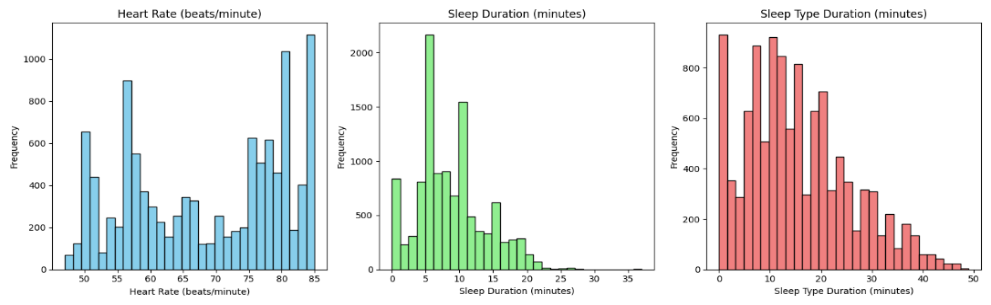
Initially, we analyzed the statistical distribution of activity data from the 100 patients. The rules for the six different types of activity data are as follows:

1. **Light Sleep Data**:
2. **No Physical Activity**:
3. **Running**:
4. **Floors Climbed**:
5. **REM Sleep**
6. **Walking**:

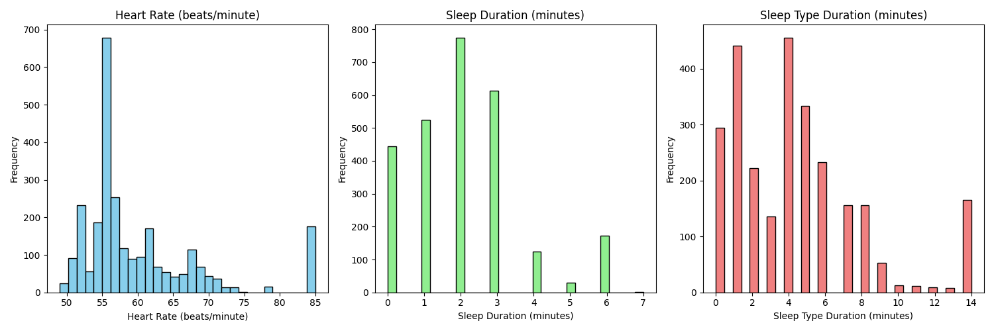


We examined the distribution of data for each activity type, focusing on light sleep, REM sleep, running, walking, no physical activity, and floors climbed.

Light Sleep Data distribution:



REM Sleep Data distribution:



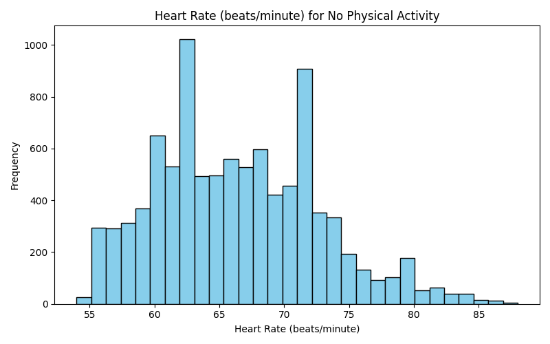
Running Data distribution:



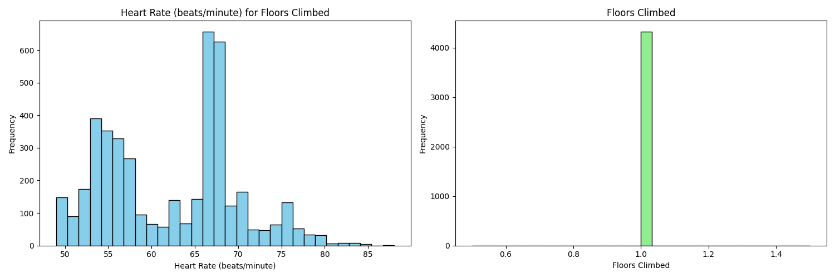
Walking Data distribution:



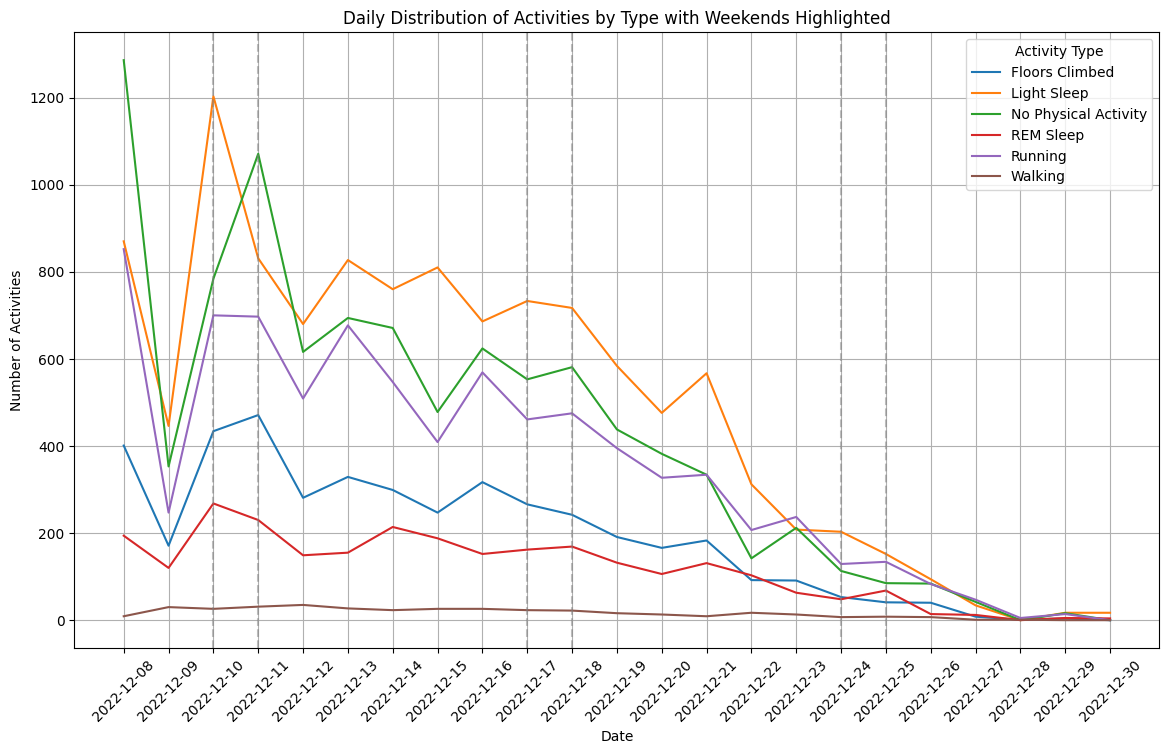
No Physical Activity:



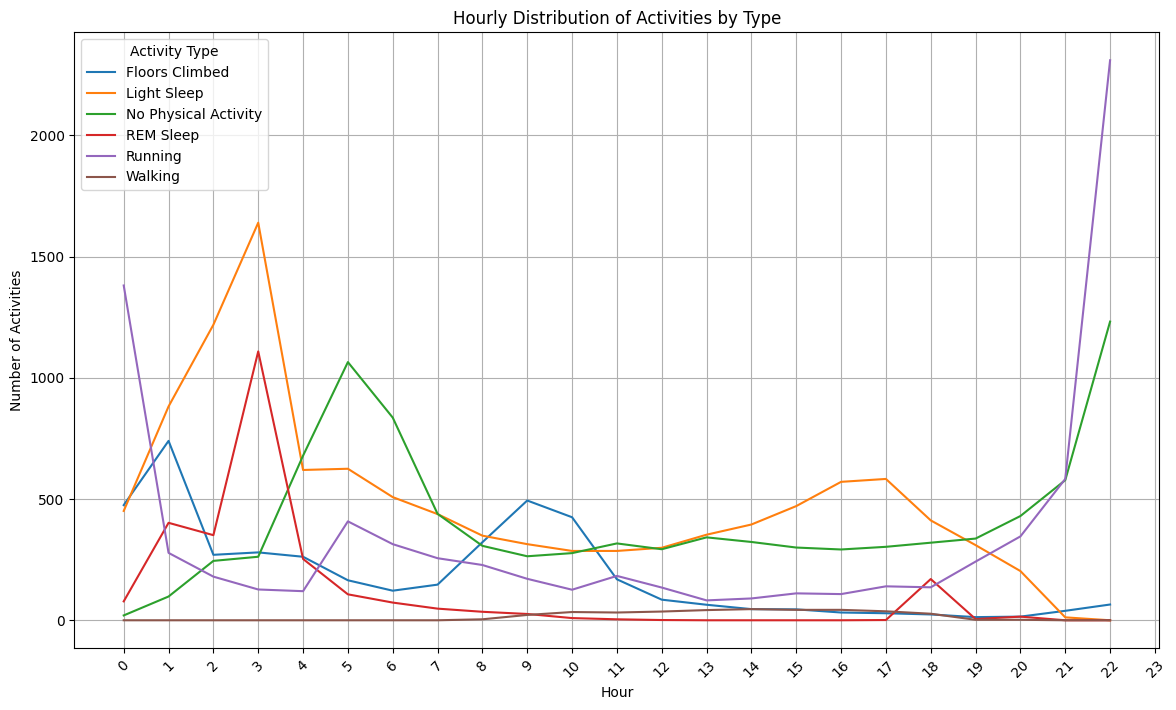
Floors Climbed:



Daily distribution of activities:



Hourly distribution of activities.



Distribution of heart rate:

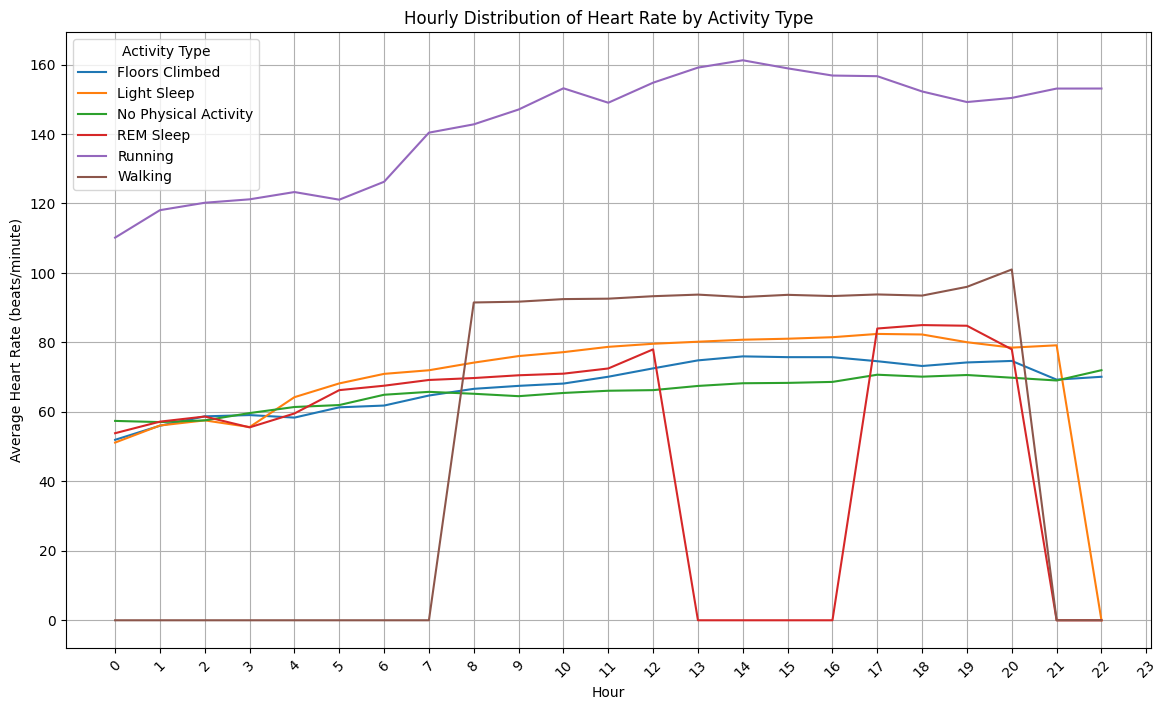


Table of hourly activity percentage distribution:

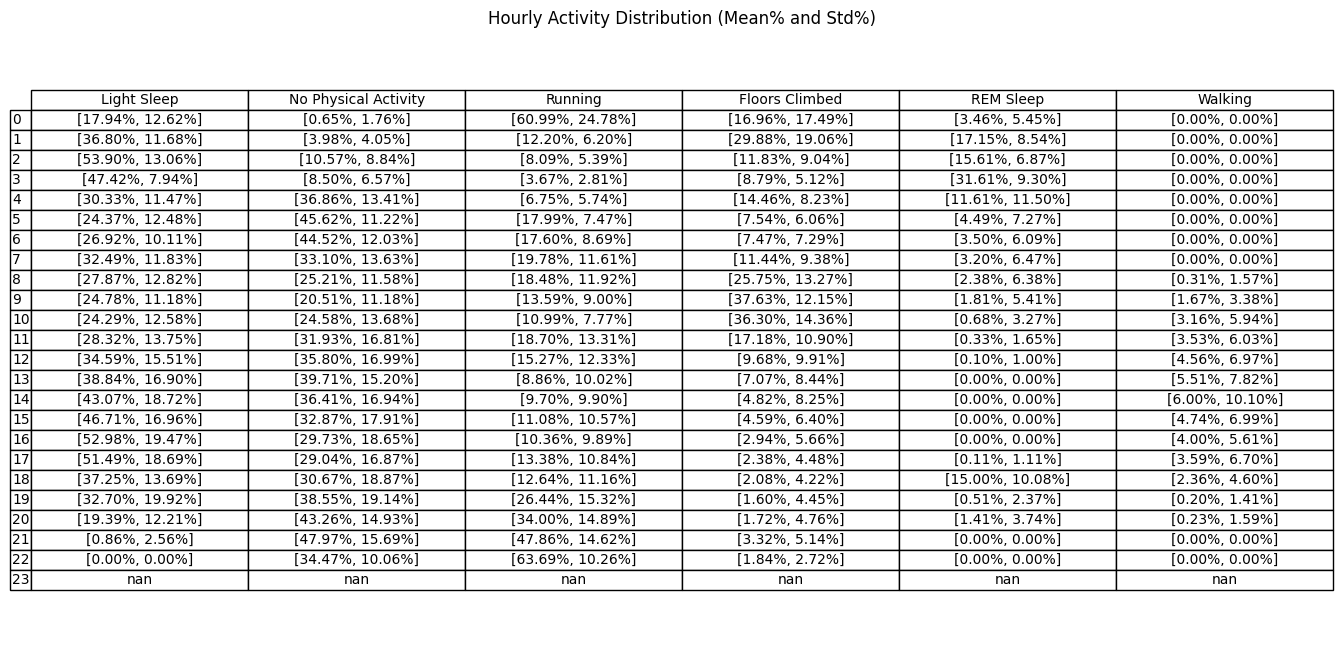
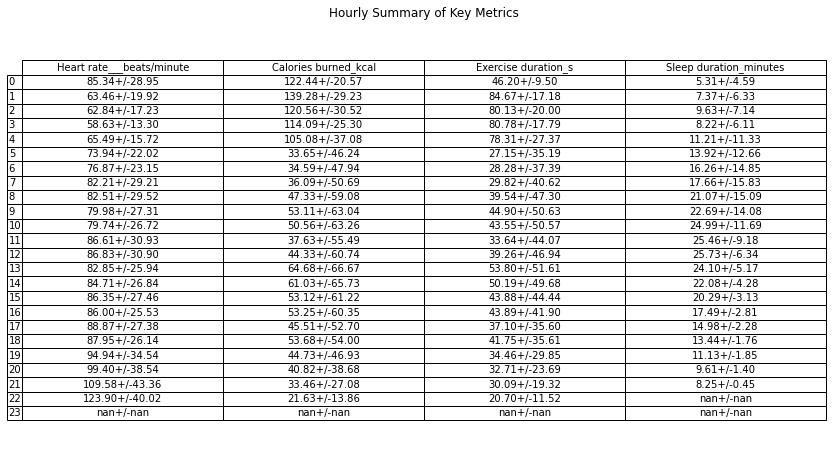
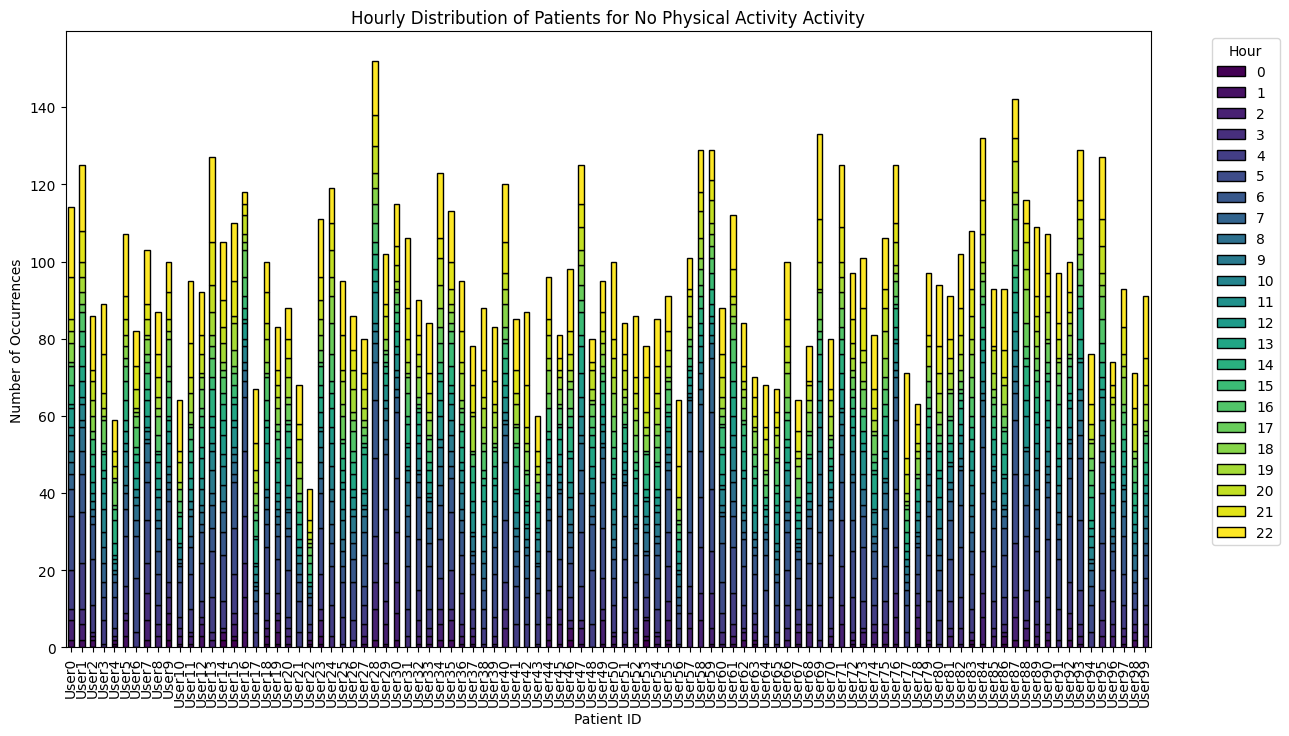
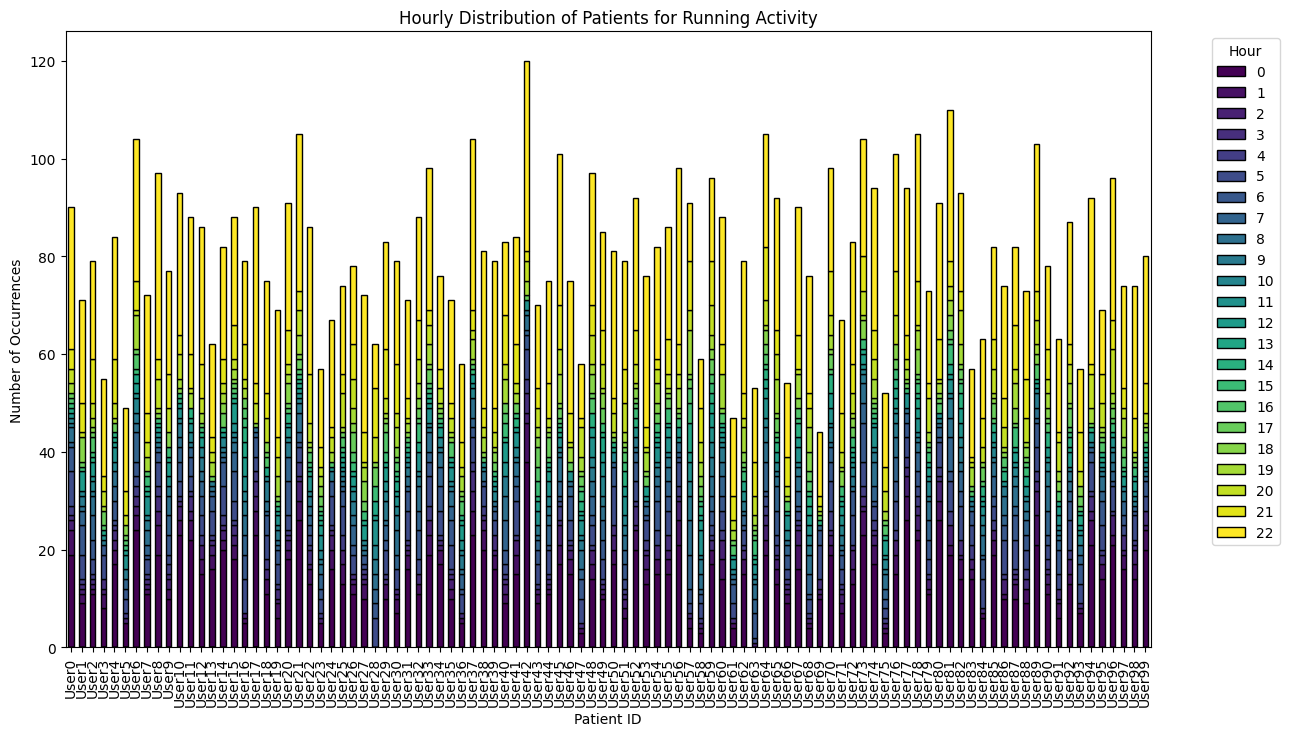


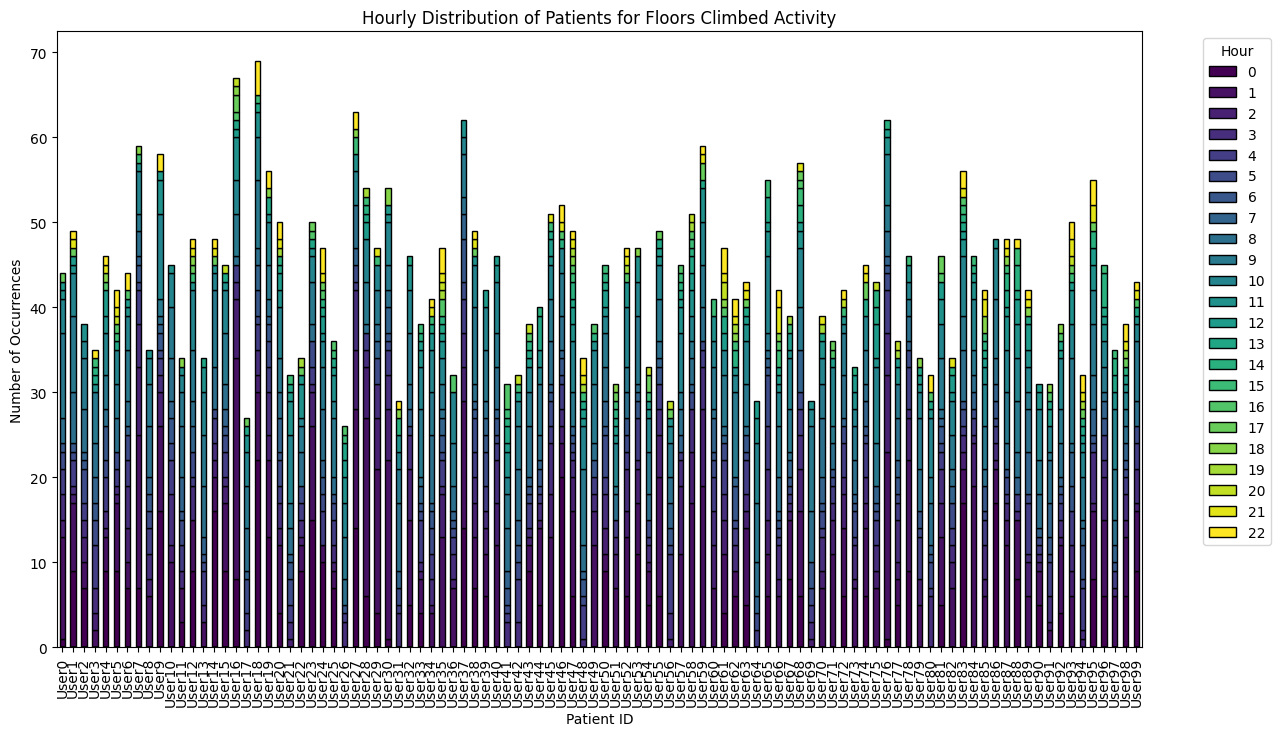
Table of hourly distribution:

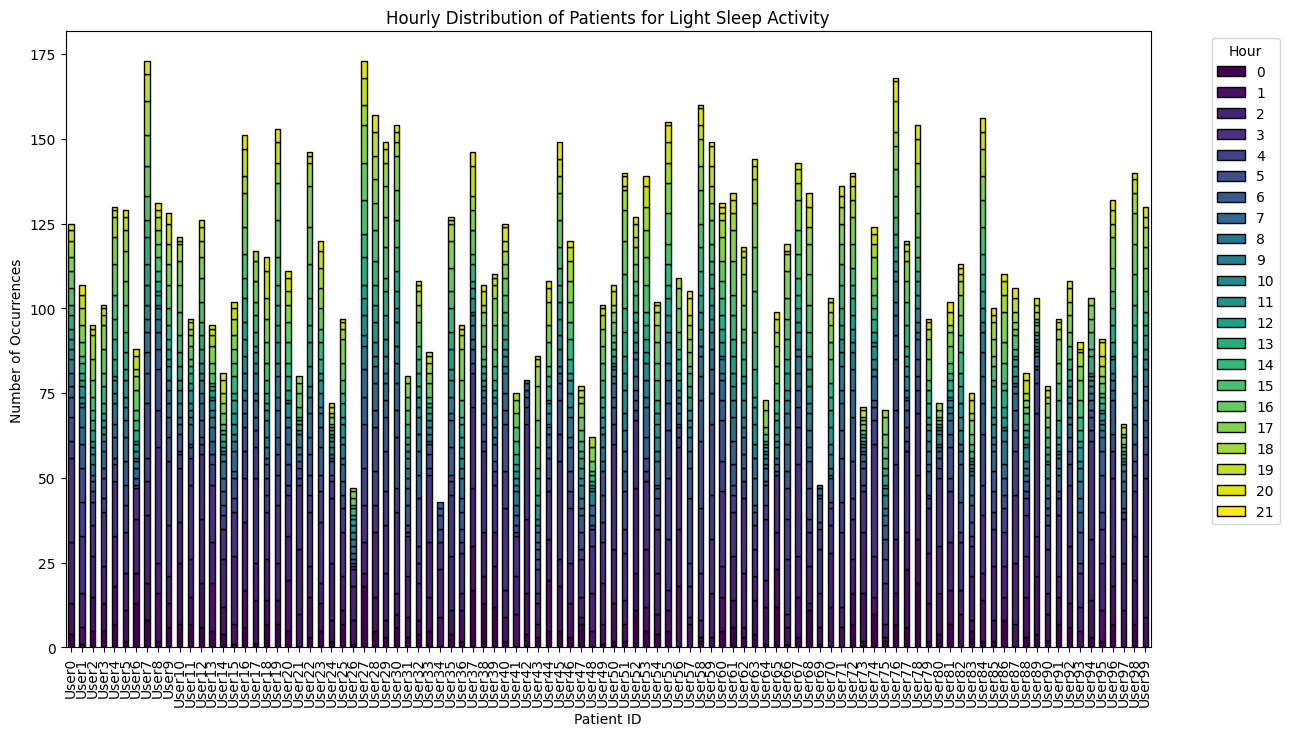


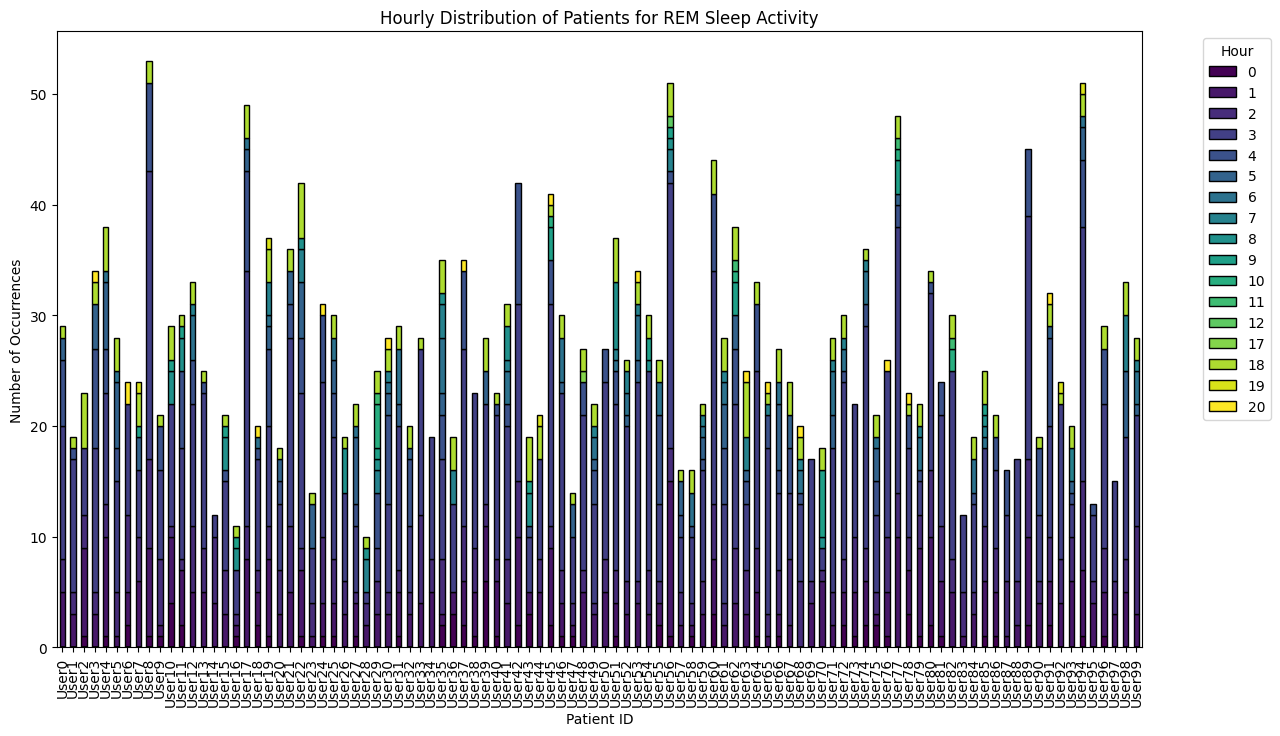
Hourly distribution of patients for different activities.

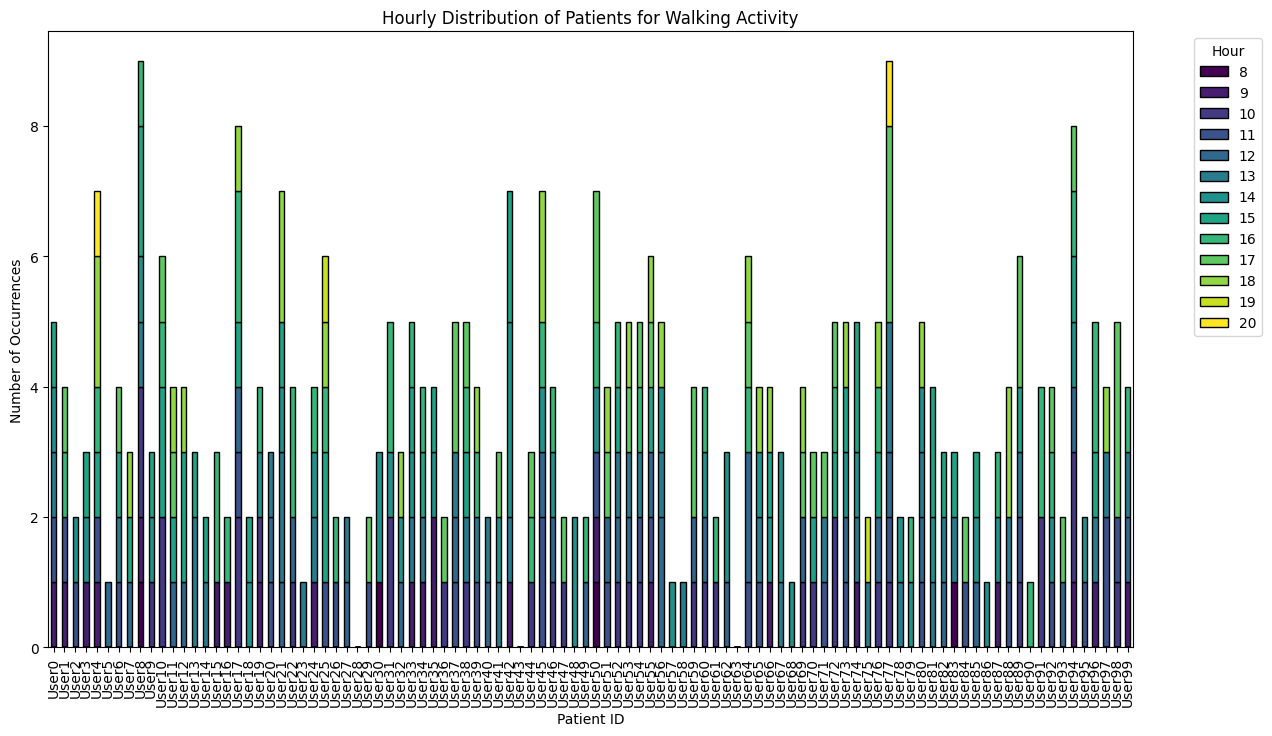












**Step 2: Synthetic Data Generation**

The synthetic data generation process involved the following steps:

1. **Datetime Generation**:
   * Generated datetime points for new patients by converting daily hours and minutes into integers (1 to 1440) and learning the distribution of activities and datetime for each day.
   * Generated hours, minutes, and activities based on the learned distribution.
2. **Activity Data Generation**:
   * Utilized the CTGAN model to generate activity data for each type, as follows:
     + For light sleep and REM sleep: generated heart rate, sleep duration, and sleep type duration.
     + For running and walking: generated heart rate, calories burned, and exercise duration.
     + For floors climbed: generated heart rate and floors climbed.
     + For no physical activity: generated heart rate.
3. **Data Compilation**:
   * Combined all generated activity data, integrating datetime and activity data to create the final synthetic dataset for new patients.

**Step 3: Synthetic Data Evaluation**

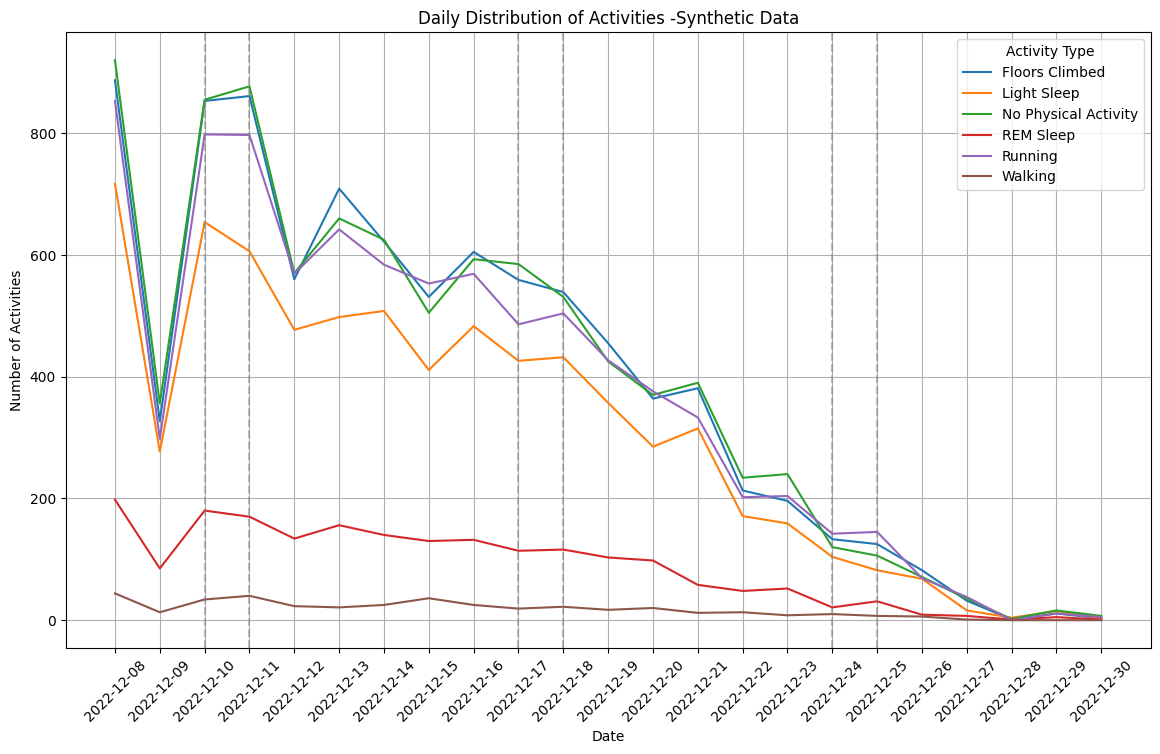
The evaluation of the synthetic data utilized several statistical metrics, including:

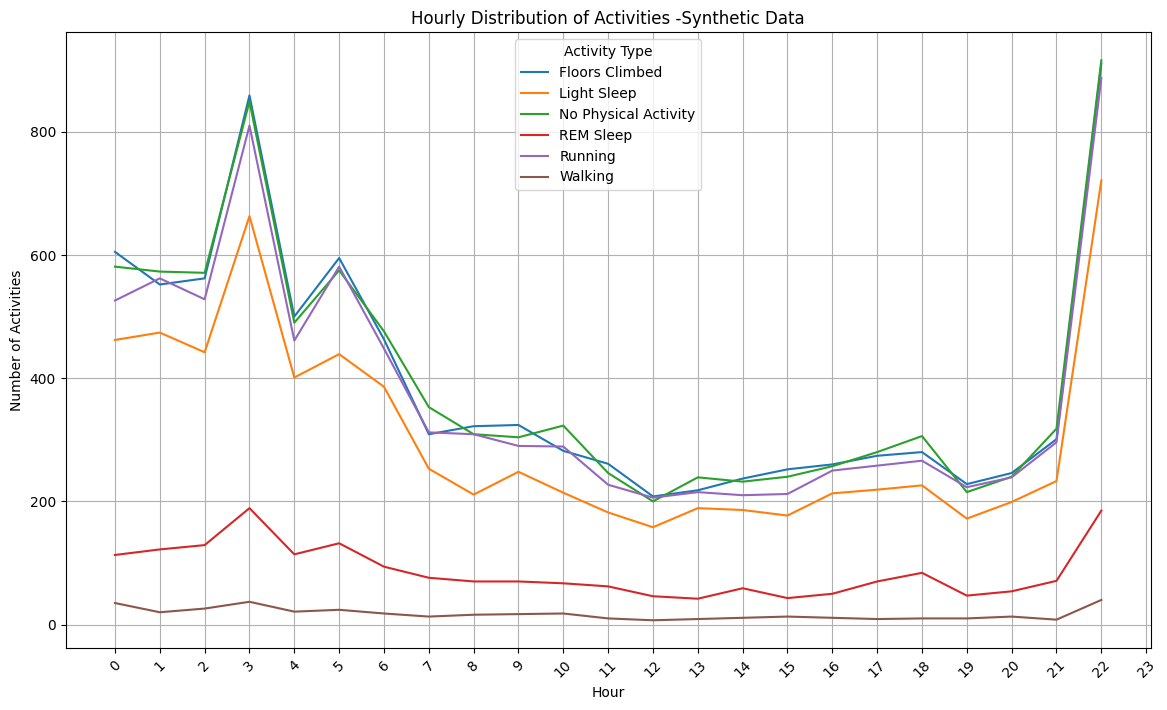
* **Wasserstein Distance**: A measure of the distance between two probability distributions.
* **Kolmogorov-Smirnov (KS) Test**: A non-parametric test for the equality of continuous distributions.
* **Jensen-Shannon Distance**: A method of measuring the similarity between two probability distributions.
* **Distance Pairwise Correlation**: Evaluating the correlation between generated and original data points.

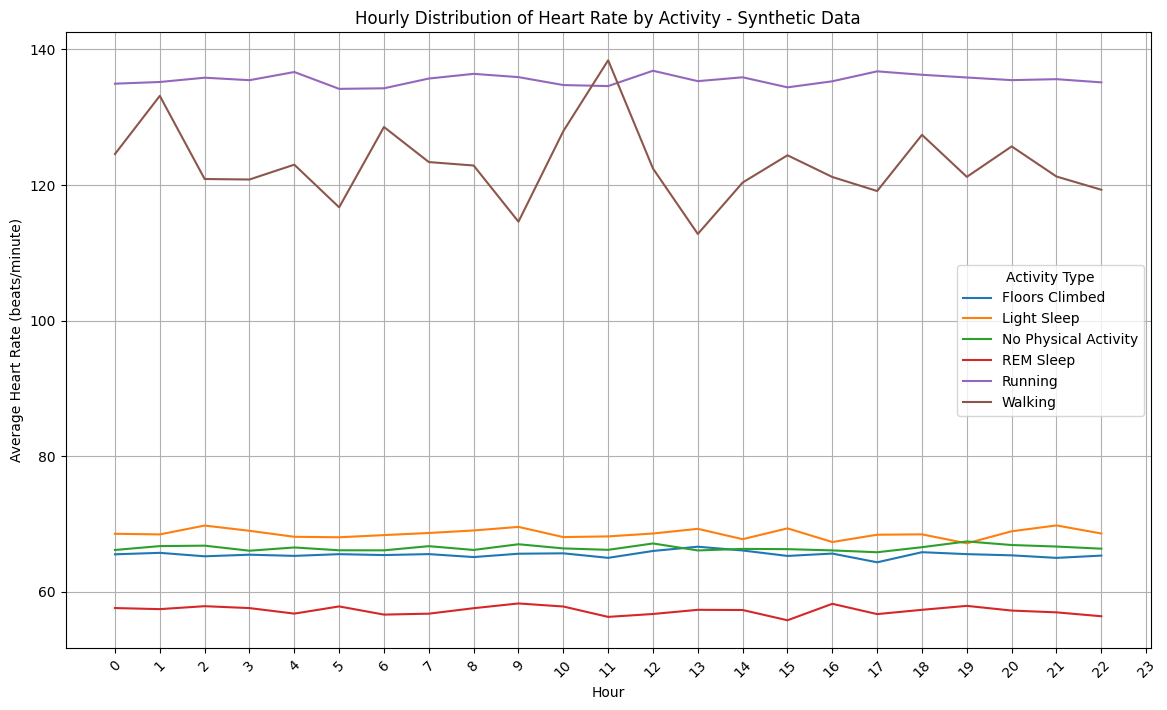
Details regarding the code for Steps 2 can be found in the accompanying zip file.

generate\_synthetic\_data.ipynb used to generate synthetic data. Generate\_User1.csv is the generated 1 patient’s synthetic data. Preliminary\_analysis.ipynb used to preprocess the original data.

**Results**







**Discussion**

Future work could explore alternative GAN [1] models, such as CTAB-GAN+ [2] and diffusion models, to further enhance the quality of synthetic data generation, focusing on learning temporal and dynamic distributions across different activities and patient profiles.

**Questions**

Reference:

1.CTGAN: Xu, Lei, et al. "Modeling tabular data using conditional gan." Advances in neural information processing systems 32 (2019).

2.CTAB-GAN+: Zhao Z, Kunar A, Birke R, Van der Scheer H, Chen LY. CTAB-GAN+: enhancing tabular data synthesis. Front Big Data. 2024 Jan 8;6:1296508.