**Synthetic Health Activity Data: Improving Data Quality and Analysis in Activity Monitoring**

Fan Dong, Wenjing Guo, Bailang Liu, Jie Liu

**Introduction**

In recent years, synthetic health activity data has become an important topic in the fields of activity recognition and health monitoring, especially as privacy concerns for real patient data. Synthetic datasets let researchers work with realistic data while keeping patient information private. In addition, Synthetic data generation helps manage missing data in health records by mimicking the statistical patterns of real data with minimal loss. This paper presents a comprehensive process for generating and evaluating synthetic activity data from 100 virtual patients, encompassing a range of activities such as light and REM sleep, running, walking, inactive states, and floors climbed. Leveraging CTGAN (Conditional Tabular Generative Adversarial Network) models, we generate a dataset that accurately mirrors the statistical distributions of actual patient activity metrics, including heart rate, calories burned, exercise duration, and sleep patterns. To validate the generated data, we conduct an extensive analysis of distribution patterns using four distinct evaluation metrics, focusing on heart rate, calorie expenditure, exercise duration, and sleep duration. This work contributes valuable insights into synthetic data generation, aiding future advancements in privacy-preserving, high-quality datasets for activity recognition and health-related research.

**Method**

**Statistical Analysis for original activity data**

We began by examining the statistical distribution of activity data from 100 patients across six types of activities: light sleep, REM sleep, running, walking, no physical activity, and floors climbed. Figure 1 presents these activity types and their distribution across patients. Key observations include: (1) heart rate is recorded for all activities, (2) calories burned and exercise duration are recorded for running and walking, (3) sleep duration metrics are recorded for light sleep and REM sleep, and (4) floors climbed are tracked only for the floors climbed activity.

Next, we analyzed each activity type’s distribution separately in Figure 2, where (a) shows light sleep, (b) REM sleep, (c) running, (d) walking, (e) no physical activity, and (f) floors climbed. To explore weekly patterns, Figure 3 shows daily activity distribution, with weekends marked by dashed lines; no consistent weekly patterns were observed across the activities.

Hourly patterns were further examined in Figure 4. We found that light and REM sleep typically occur in the early morning and late afternoon, while other activities showed no consistent hourly trends. In Figure 5, heart rate distributions are compared by activity: running had the highest average heart rate around 140 bpm, followed by walking at 90 bpm, with other activities near 70 bpm. Figure 6 aggregates activities into time periods—night (0–6 am), morning (6 am–12 pm), afternoon (12–6 pm), and evening (6 pm–12 am)—showing the highest activity at night, followed by afternoon, morning, and evening. Figure 7 highlights activity trends in these time periods, revealing that light and REM sleep occur most at night, running in the evening, walking in the afternoon, floors climbed in the evening, and no physical activity evenly across periods.

Table 1 shows hourly activity distribution percentages, offering guidance for generating synthetic data that mirrors activity patterns across time.

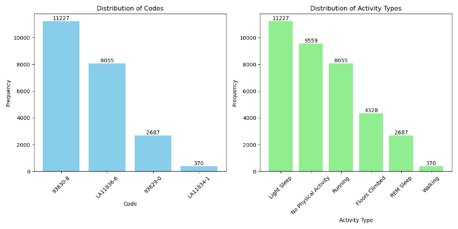


Figure 1. Activity distribution for all patients

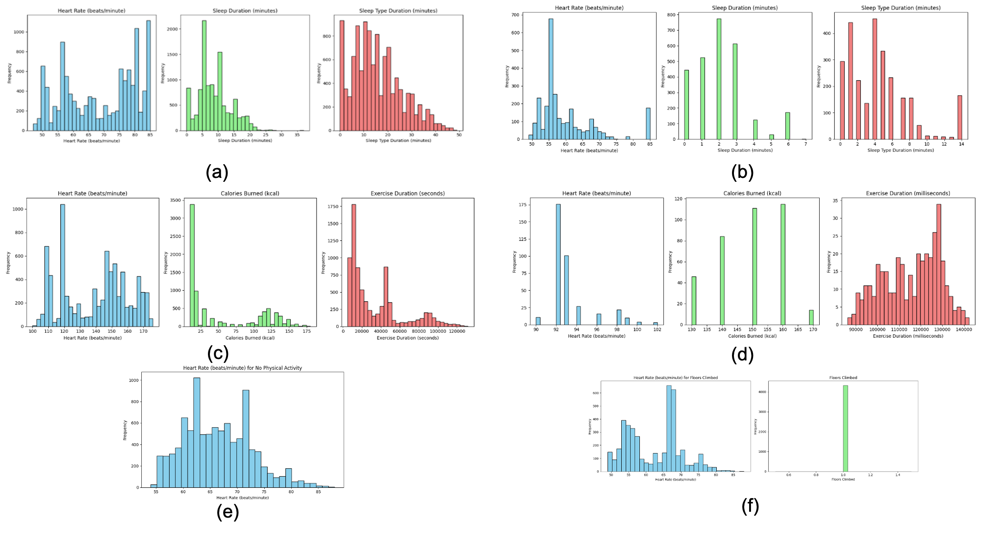


Figure 2. heart rate, sleep duration, calories burned, floor climbed distribution for different type of activities.

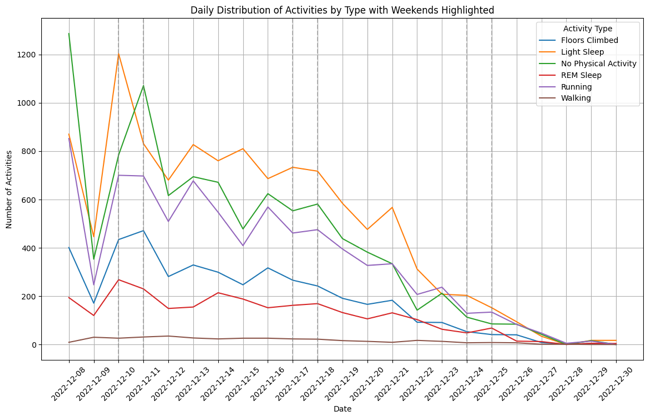


Figure 3. Daily activity distribution for original data

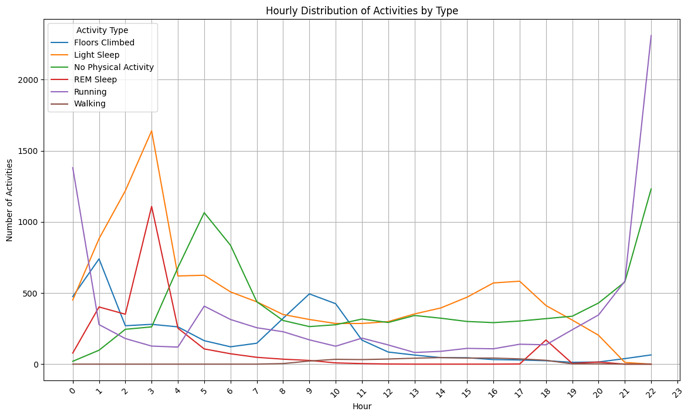


Figure 4. Hourly distribution of activities for original data

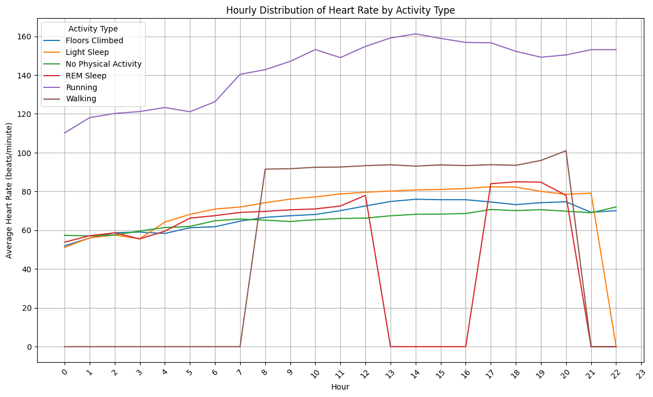


Figure 5. Distribution of heart rate for different activities for original data.

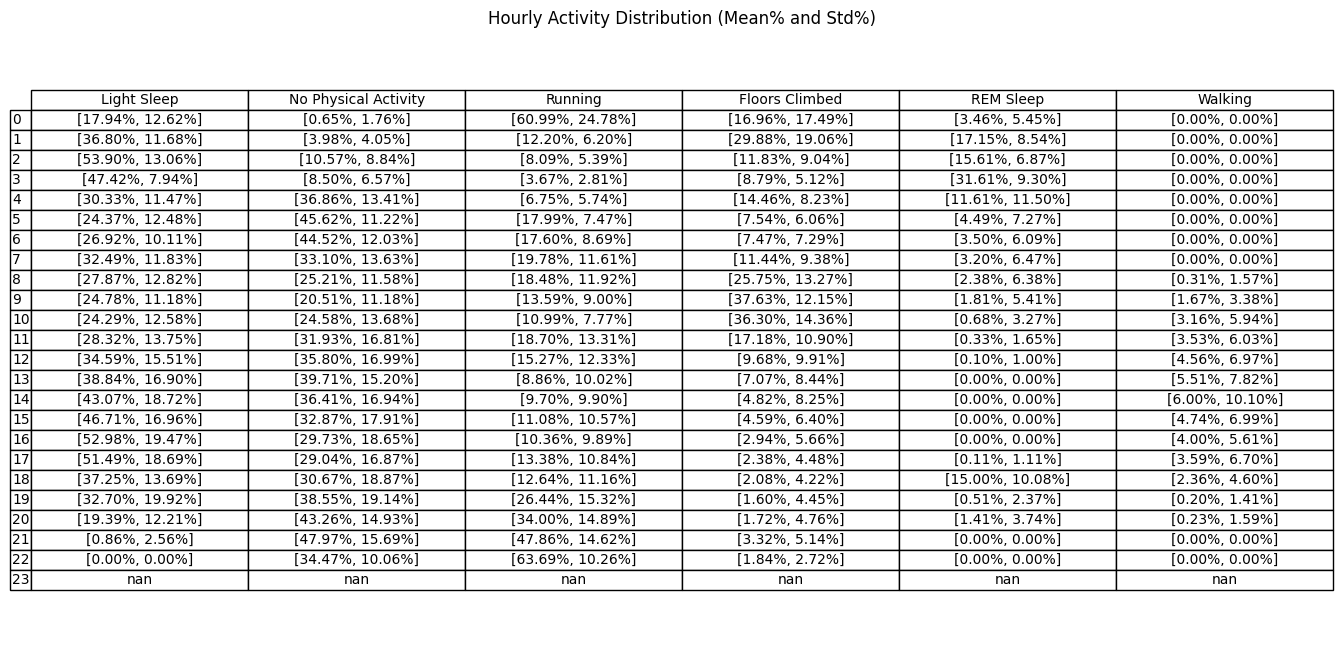


Table 1. Table of hourly activity distribution.

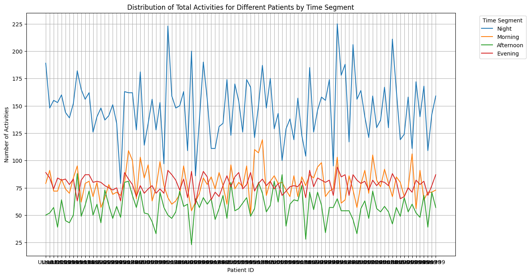


Figure 6. aggregate activities in different time period.

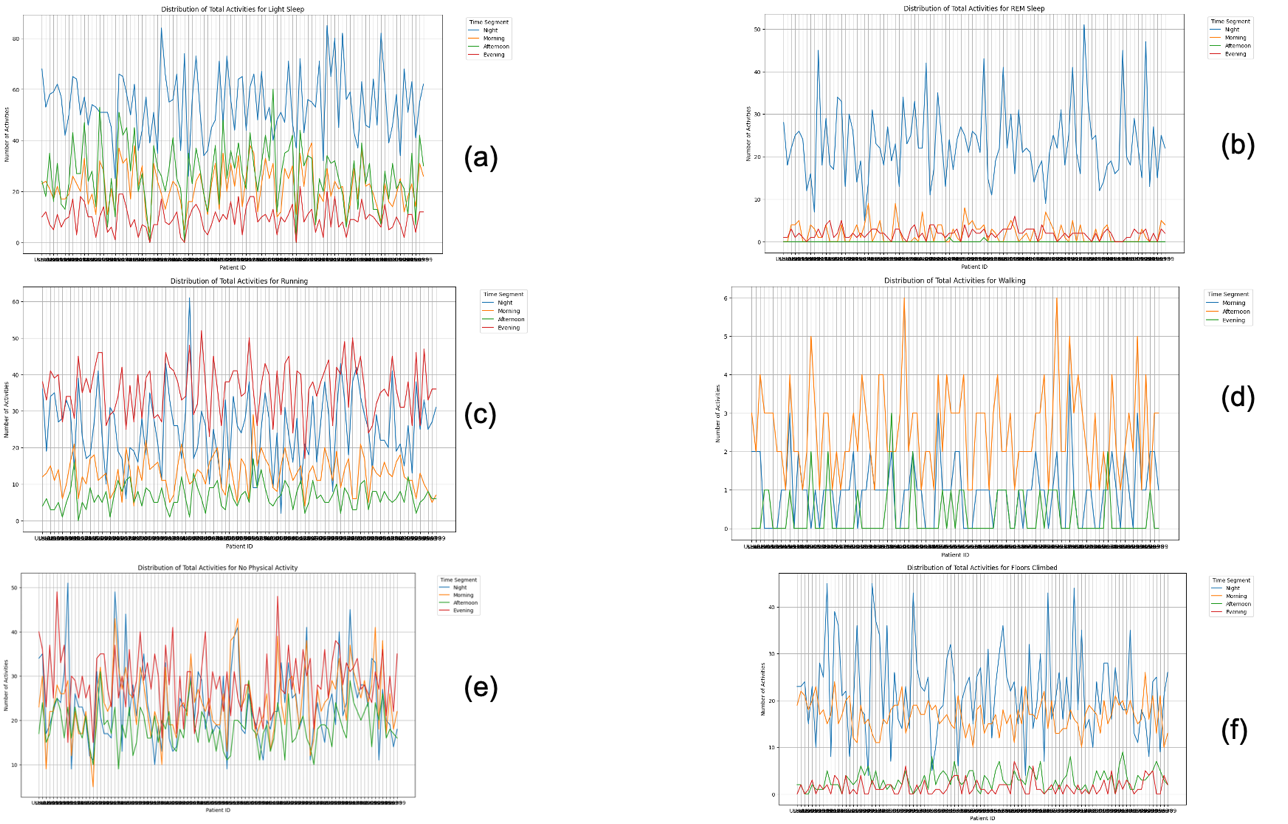


Figure 7. different activities in different time period.

**Synthetic Data Generation**

The synthetic data generation process for creating realistic patient activity data involved three key steps:

1. **Datetime Generation**:
   * The start datetime was set as 2022-12-08 00:00, encoded as 1, and the end datetime as 2022-12-30 23:59, encoded as 33120, covering a total duration of 23 days (23 days × 24 hours × 60 minutes). A distribution of activities and timestamps was learned across all patients to capture typical activity patterns, and this learned distribution was then used to generate datetime points for 100 new synthetic patients. These generated datetime integers were later converted into timestamps.
   * Based on Table 1, we generated different activity types for each synthetic patient, ensuring that hourly distributions adhered to the constraints observed in the original data.
2. **Activity Data Generation**:
   * Using the CTGAN model, activity data was generated for each activity type, capturing realistic values for essential metrics:
     + 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**:
   * The generated timestamps and activity types from Step 1 were combined with the detailed activity data (heart rate, calories, sleep duration, and floors climbed) from Step 2 to create the final synthetic dataset, representing a new population of 100 patients with realistic activity patterns.

**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.

**Results**

Figures 8, 9, and 10 can be compared with Figures 3, 4, and 5, respectively, showing the distribution of original and synthetic data. These figures reveal similar patterns between the two datasets.

Table 2 presents the mean and standard deviation for heart rate, calories burned, exercise duration, and sleep duration. It highlights that during evening activities from 12 AM to 4 AM, the original data shows high calorie expenditure (over 130 kcal) and low heart rates (around 65 bpm). In contrast, the synthetic data has lower calories burned and heart rates. This discrepancy may be due to a higher percentage of running activities occurring in the early morning in the original data, which is not reflected in the synthetic data. Overall, heart rate and sleep duration in the synthetic data are more closely aligned with the original data than calories burned and exercise duration.

Table 3 evaluates four metrics comparing synthetic data to the original data. The Wasserstein Distance indicates that heart rate and sleep duration are more accurately simulated in the synthetic data than calories burned and exercise duration, which show higher divergence. The highest Wasserstein distance is found for calories burned (4.6976), revealing significant differences in both the Wasserstein and KS test metrics, suggesting that the synthetic data struggles to represent this aspect accurately.

Heart rate appears well-preserved regarding distribution and relationships, as indicated by lower Wasserstein and Jensen-Shannon distances. Exercise duration shows a moderate difference, but it is less than that of calories burned, indicating reasonable fidelity to the original data. The Distance Pairwise Correlation suggests that while some relationships are maintained, the structure of the synthetic data differs noticeably from the original, indicating a need for further adjustments to enhance realism.

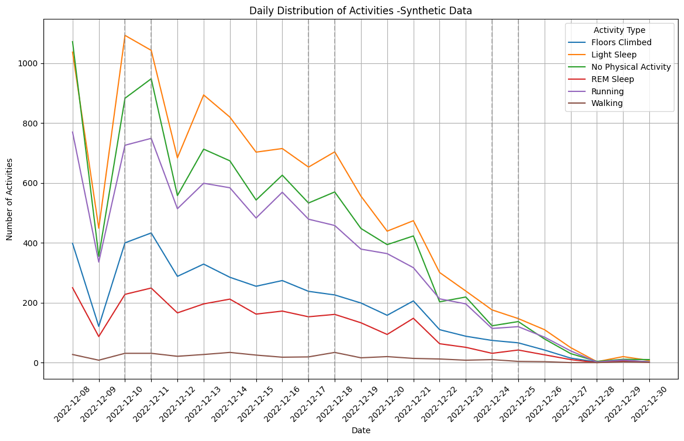


Figure 8. Daily activity distribution for synthetic data

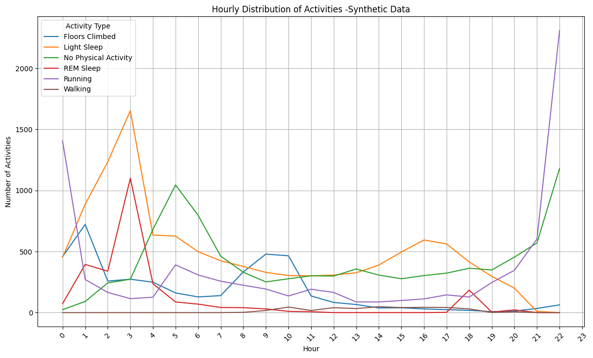


Figure 9. Hourly distribution of activities for synthetic data

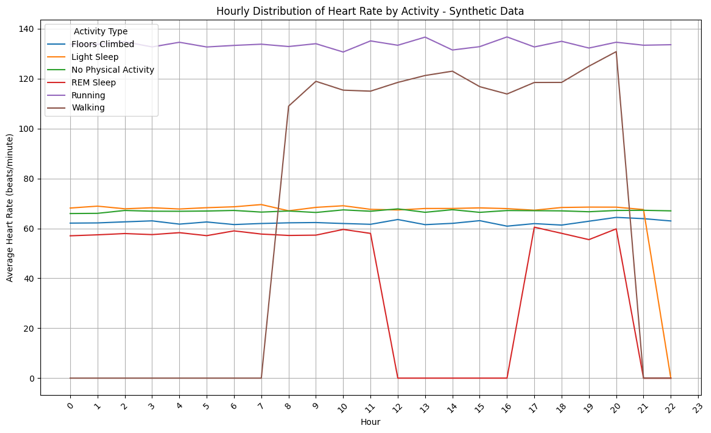


Figure 10. Distribution of heart rate for different activities for synthetic data

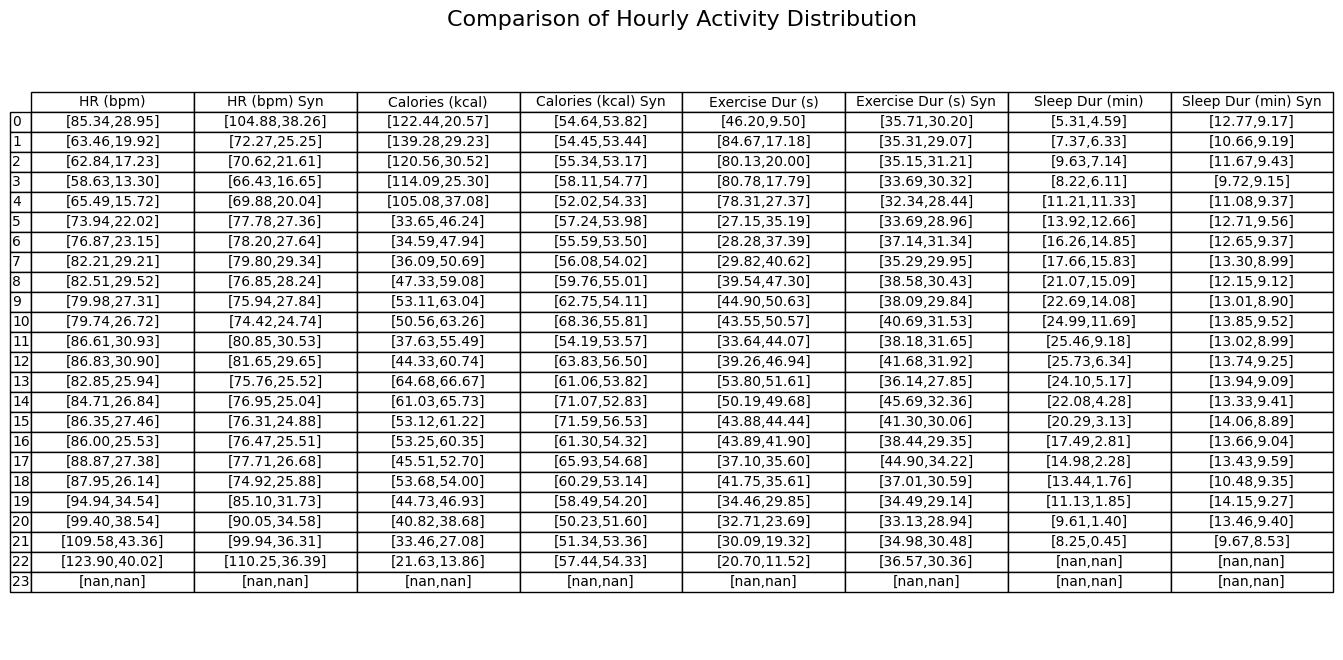


table 2. Comparing of hourly activity detail heart rate, calories, exercise duration and sleep duration between original data and synthetic data.

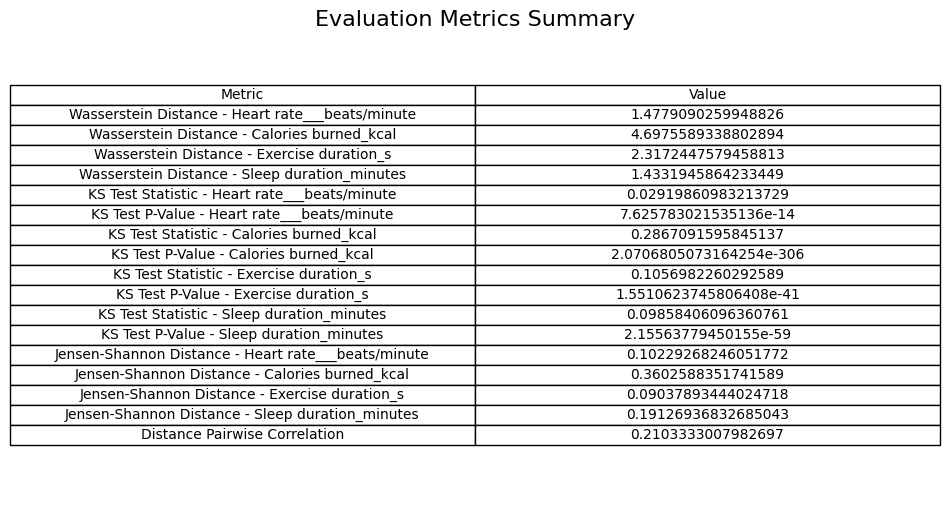


Table 3. Final Evaluation metrics for synthetic data

**Discussion**

The related code and data could be found : <https://github.com/FANMISUA/BHI_Track1>.

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.

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.