Daily Activity Analysis with Fitbit Tracker Data

Jakov Basic

December 2023

1. Introduction

Based on current statistics, there is a concerning public health issue with adult physical activity levels. According to the World Health Organization's Global Status Report on Physical Activity 2022, more than 80% of adolescents and 27% of adults do not meet the recommended levels of physical activity. This lack of physical activity not only affects the health of the individual, but also places an economic burden on health services and society as a whole. [1]

Given this context, wearable technologies like Fitbit can play a crucial role in addressing the challenges of physical inactivity. Fitbit and similar devices are designed to track various aspects of physical activity, including steps taken, calories burned, heart rate, and even sleep patterns. This data can be extremely valuable for users in several ways.

By providing real-time feedback on activity levels, Fitbit encourages people to be active and set realistic fitness goals. This can be particularly motivating for those who are not naturally inclined to exercise. The data collected can help users better understand their own exercise habits and make informed decisions about their health and fitness routines. Regular use of Fitbit can help people adopt healthier lifestyles. For example, device reminders to move or exercise can help users reduce sedentary behavior and increase physical activity in their daily lives. [2]

For healthcare professionals, data from Fitbits can provide valuable insights into a patient's daily activity levels, which can be used to tailor health advice and interventions more effectively. On a larger scale, aggregated data from wearable devices can help researchers and policymakers understand population-level trends in physical activity, informing public health strategies and interventions. [3]

The objective of this study is to explore the intricate relationships between various activity parameters as recorded by wearable technologies, specifically Fitbit trackers. The intent is to align our investigation with the World Health Organization’s efforts to enhance physical activity levels globally, addressing the alarming trend of physical inactivity among adults and adolescents.

Data collected from Fitbit devices will be analyzed to understand the patterns of physical activity among various demographics. A key goal is to identify correlations between different activity parameters. Another goal is to gain insights about behavioral trends related to physical activity. This involves understanding what motivates individuals to be more active and identifying barriers to physical activity. The objective and methods will be discussed in more detail in the following chapters.

2. Problem formulation

The primary issue of the project is to explore and understand the complex relationships and interactions that exist between different forms of daily physical activity. These activities fall into three different types:

1. **Active**: this includes time with high and moderate levels of physical activity.
2. **Lightly active**: this includes time with low levels of physical activity.
3. **Sedentary**: This category includes time with minimal or no physical activity.

The aim of this project is to answer 2 questions related to these activity types:

- How does total activity throughout the day affect calories burned, and how much distance needs to be traveled to achieve a certain level of calories burned?

- How does the activity distribute throughout the week on group level and do they match any health recommendations?

By exploring these issues, the project aims to provide valuable information on effective physical activity strategies to promote health and fitness. The study will investigate how different intensities and durations of activity affect the total calories burned and the distance needed to reach specific calorie targets. This research will help to understand whether short, high-intensity exercises are more effective in burning calories than longer, lower-intensity exercises. The results will help us to gain a deeper understanding of patterns of daily physical activity and their impact on health metrics.

3. Dataset Description

Dataset is fetched from: https://www.kaggle.com/arashnic/fitbit

Daily Activity Dataset (dailyActivity_merged.csv)

The dataset provides a comprehensive overview of daily physical activities, including steps, distances, active minutes, and calorie consumption. Activity variables are continuous and numerical, and can be compared between individuals and dates. Variables are self explanatory.

- Observations: 940 entries, 33 different unique Id:s
- Variables: Id, ActivityDate, TotalSteps, TotalDistance, TrackerDistance, LoggedActivitiesDistance, VeryActiveDistance, ModeratelyActiveDistance, LightActiveDistance, SedentaryActiveDistance, VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes, SedentaryMinutes, Calories.
- Missing Data: No missing data observed in any of the columns.

Hourly Intensity Dataset (hourlyIntensities_merged.csv)

Each row in the dataset represents an hour's worth of activity data for a given individual. The dataset could be used to analyze patterns in physical activity, such as the intensity of activities at different times of the day or on different dates.

- Observations: 22,099 entries, 33 different unique Id:s
- Variables: Id, ActivityHour, TotalIntensity, AverageIntensity
- Missing Data: No missing data observed in any of the columns.

Hourly Calories Dataset (hourlyCalories_merged.csv)

Similar to the hourly intensity database, but instead of intensity variables there is only one variable describing calories burned during the hour.

- Observations: 22,099 entries, 33 different unique Id:s
- Variables: Id, ActivityHour, Calories
- Missing Data: No missing data observed in any of the columns.

Datasets seem to be well-structured and comprehensive, covering a wide array of variables pertinent to physical activity. The absence of missing data in these datasets simplifies the initial stages of data preprocessing. The datasets link by the Id column, suggesting that they pertain to the same group of individuals. All datasets can also be linked by date, and datasets containing hourly data can be linked by the hour of the day.

Data will be preprocessed to prepare the datasets for analysis. Data preprocessing is a crucial phase in any data analysis project as it involves cleaning, transforming, and organizing raw data into a suitable format for analysis. This process ensures the quality and integrity of the data, facilitating accurate and reliable outcomes.

The three primary datasets: dailyActivity_merged.csv, hourlyIntensities_merged.csv and hourlyCalories_merged.csv, will be briefly described, highlighting the significance and the types of variables they contain.

Data Cleaning

Initially, a check for missing values in all datasets was performed. It was observed that there were no missing values, simplifying the data cleaning process.

Assessing Relevancy:

Some of the data was immediately dropped, since it was considered irrelevant considering the objective. Reasons for dropping the values:

- TrackerDistance: Very high correlation and similar values with TotalDistance, but less trustworthy. No need for 2 nearly identical variables.
- LoggedActivitiesDistance: Irrelevant and already covered by TotalDistance
- VeryActiveDistance and ModeratelyActiveDistance: Merged with each other.
- SedentaryActiveDistance: values near 0 for almost every entry.
- VeryActiveMinutes and FairlyActiveMinutes: Merged

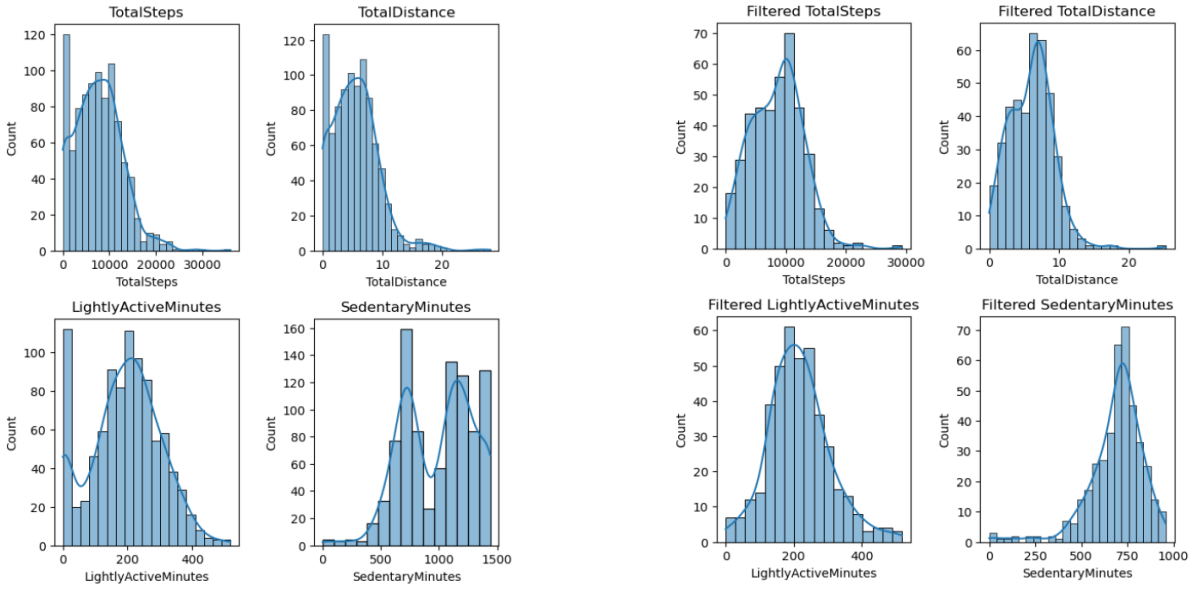
Dealing with Outliers and Invalid Data:

High frequency of 0 min activity and distance values, combined with near to 24h sedentary time raised suspicion. Entries with abnormally high sedentary time might indicate misuse of the tracker or data recording errors, thus compromising data integrity.

Since a sedentary period exceeding 16 hours in a day may not accurately represent a typical day's activity pattern for most individuals, entries with over 960 min ‘SedentaryMinutes’ were excluded.

After excluding the entries affected by the probable misuse, Interquartile Range (IQR) method was used for removing outliers from the data. Outliers can significantly skew the results of the analysis. Using visual methods like hist and box plots, potential outliers in variables were identified and assessed.

The changes in distributions of certain variables can be seen from the histograms below. Since the pattern is similar to every variable, I have not included all of them in this visualization:



As we can see, the 0 values in step, distance and activity affect the shape of the distribution remarkably. We cannot know for certain which values are false, but we have made an assumption that values with high sedentary and zero activity time are caused by misuse of the tracker.

Data Consistency Checks:

Ensured consistency in data formats, especially dates and times were converted to datetime format, in all datasets. This step is critical for merging datasets and conducting time-based analyses.The data was also checked for any obvious errors or inconsistencies, such as negative values in steps or calories, which are not feasible in the given context.

Data Transformation

Feature Engineering:

New features were derived from existing data to provide more insights, and clarify the visualization. Since ‘VeryActiveMinutes’ and ‘FairlyActiveMinutes’ both indicate increased physical activity, they were combined to ‘ActiveMinutes’. It also made visualization more compact.

‘Weekday’ column was created by transforming ‘ActivityDate’ into names of weekdays to help visualize typical behavior for each weekday.

Normalization and Scaling:

To bring numerical variables onto a comparable scale, I used Min-Max scaling.

4. Methods

Exploratory Data Analysis

Exploratory data analysis is an essential step in data science, allowing us to discover patterns, spot anomalies, frame hypotheses, and check assumptions through summary statistics and graphical representations.

Descriptive Statistics:

Measures of central tendency include the mean, which calculates the average value for variables, and the median, which identifies the middle value in ordered data, making it valuable for handling skewed data.

Measures of spread, such as the standard deviation, were used to measure the extent of variation or dispersion within a dataset. High standard deviation suggests that the data points are more widely spread out.

The ‘corr’ method was used to calculate the correlation coefficient, specifically the Pearson correlation coefficient, to assess the relationship between various variables in our dataset.

Data Visualization

Histograms and box plots were utilized to visualize the distribution of each variable, providing insights into distribution shapes and highlighting outliers and data spread.

Bar and line plots were used to examine the behavior of activity variables over the week. These plotting methods are commonly used for time series, and serve the purpose well.

For a more detailed behavioral examination I used heatmaps. Heatmaps are commonly used when examining intensities between hours for multiple days.

Heatmaps were also used for identifying overall correlations in the dataset. These can be used to visually represent the correlation matrix, making it easier to identify which variables are most strongly correlated. However it is not relevant to present the correlation matrix heatmap in this project, since we have more detailed visualization methods for the variables we want to examine.

More detailed relationships of the variables were visualized with scatter plots. Scatter plots are ideal for visualizing relationships between two variables.

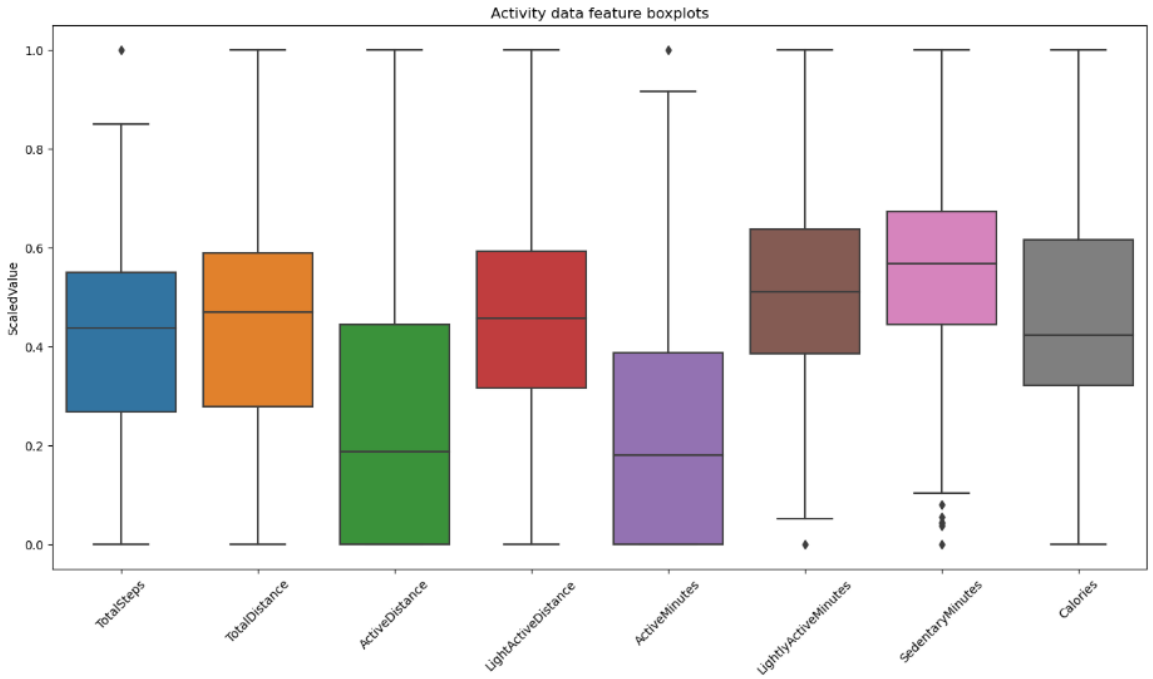
5. Results

We began by examining the basic statistical measures of our data, such as mean, median, and standard deviation. These measures provided a foundational understanding of the typical values and variabilities within our key metrics.

Here we have the filtered data. Even after we filtered a lot of entries having high sedentary minutes, we still have min value of 0 for many activity variables.

	TotalSteps	TotalDistance	ActiveDistance	LightActiveDistance	ActiveMinutes	LightlyActiveMinutes	SedentaryMinutes	Calories
count	371.000000	371.000000	371.000000	371.000000	371.000000	371.000000	371.000000	371.000000
mean	8367.137466	5.869164	2.078221	3.783801	39.867925	214.253369	716.137466	2327.916442
std	3819.654618	2.715012	2.113022	1.534423	40.343007	75.865389	104.972274	693.696067
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	413.000000	665.000000
50%	8758.000000	6.220000	1.570000	3.780000	30.000000	210.000000	722.000000	2175.000000
max	20031.000000	13.240000	8.430000	8.270000	167.000000	412.000000	957.000000	4236.000000

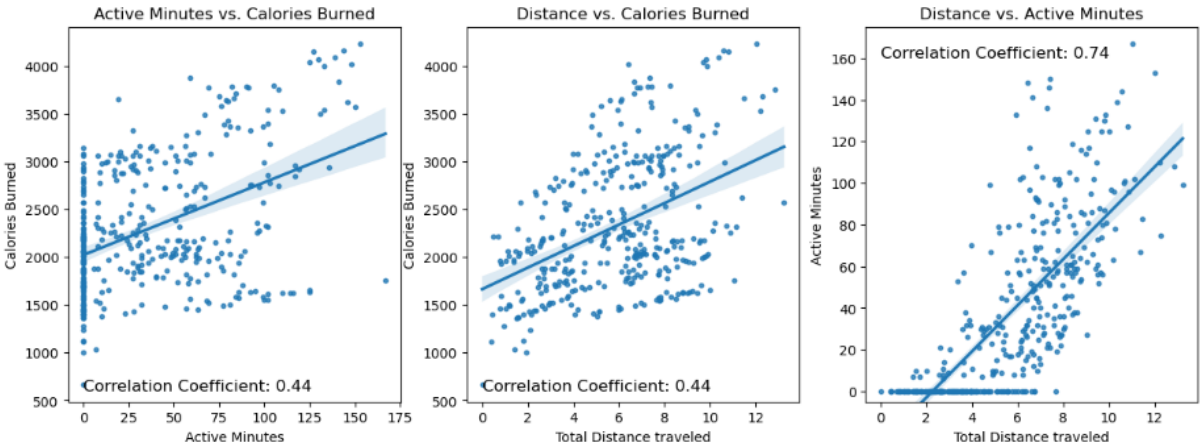
As most of the data are not comparable with each other, we can gain more insight from scaled data. I have visualized the scaled data by using boxplot.



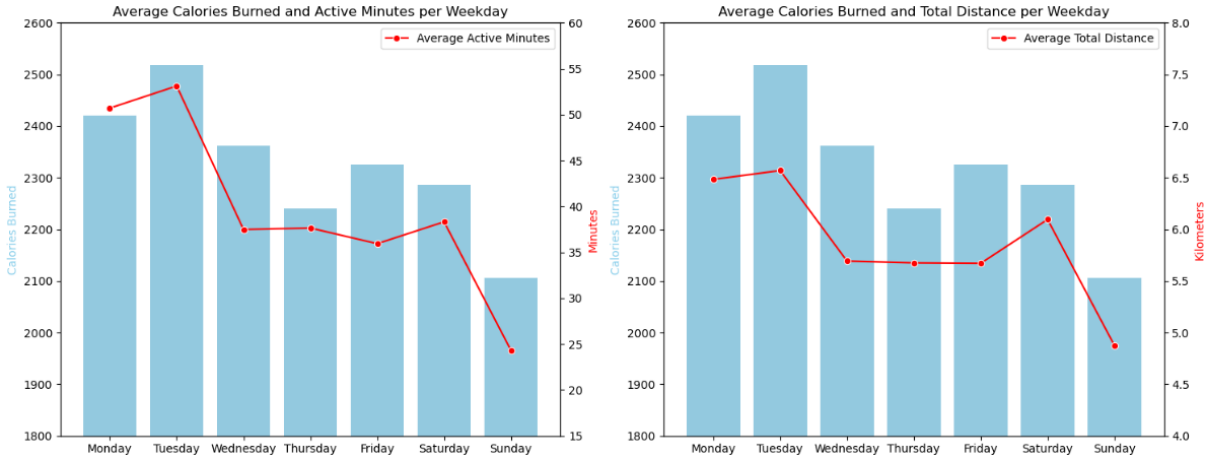
We can see that the lower quantile of the box only touches 0 for the "ActiveDistance" and "ActiveMinutes" values. I consider the data to be valid, as not everyone

exercises every day. Also the relatively high scaled median of sedentary minutes could cause low activity levels.

The main goal of this project was to investigate how different activity types affect burned calories and how much distance traveled it equals. The correlation matrix showed that both ‘ActiveMinutes’ and ‘TotalDistance correlate relatively well with ‘Calories’. To investigate the correlations of these variables further, I decided to use scatter plots with fitted regression lines.



We can see from the plots above, that the chosen variables have some positive correlation. To see how these variables behave over time, I conducted summarized time series analysis, which shows group level averages for different weekdays. The aim was to see if there was a pattern in activity over the week and how well the active time and distance traveled affected the calories burned. Since the correlation was relatively high, I expected the variables to follow a pattern similar to each other.



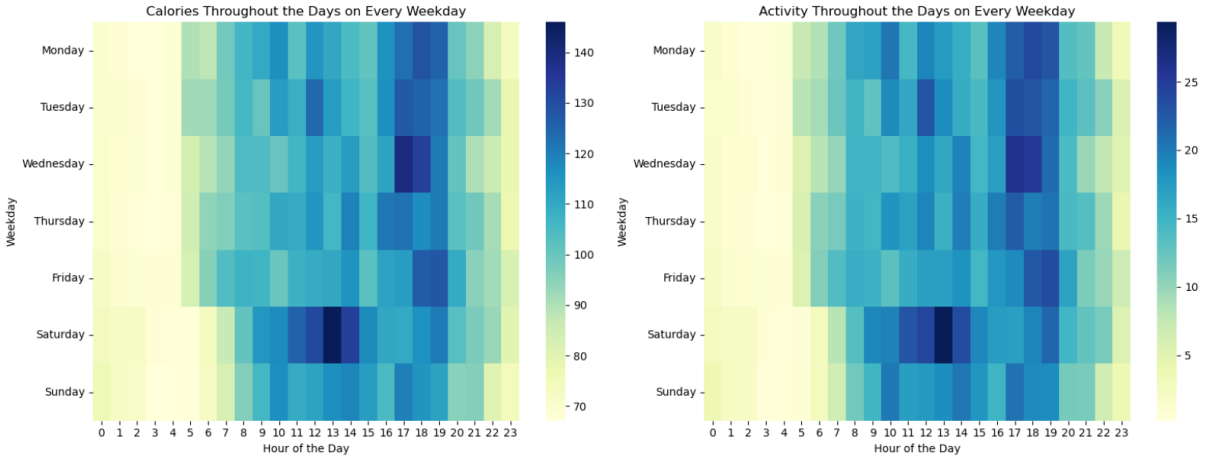
As we can see from the plot above, the plots indeed follow a similar pattern. To improve comparability I could use scaled values, but showing the original values serve a purpose in understanding how the values of activity time and distance

traveled correspond to the burned calories. However, it is important to remember that these are just average values. For example, the twin axis plot with calories and distance shows that burning 2400 calories a day, corresponds to 7 kilometers of traveling. If we look at the scatter plot that has Distance vs. Calories Burned, we can see that when the x-axis equals 7 Km, the entries on the y-axis are spread with a range approximately from 1400 to 3800 Calories.

The plot shows clear differences in the data on different weekdays. The average burned calories per day varies from 2100 to 2550. The maintenance calories for most adult women ranges from 1600 to 2200 calories a day, and from 2200 to 3000 for adult men. We do not have any information about the age and gender, but as the average burned calories are similar to the lower half of male maintenance calories, the data is likely gathered from an adult group that on average burns the amount of calories they use.

All the observed variables show that activity is highest in the beginning of the week. As we move into the middle of the week, the activity drops, and on Sunday, it hits its lowest point. To make sure that the information is valid, I checked the count of the entries for each weekday. There was no significant variation, so I decided not to process the data any further.

To find out how the activity is distributed throughout the days, I used heatmaps for hourly calories and intensities. Intensity is not equal to the active minutes, but it has a strong correlation and enhances the visualization.



We can see that the heatmap is almost identical for burned calories and activity intensities. The activity starts earlier during the week than on the weekends, but tends to end at the same time for the entire week, except for Sundays and Wednesdays. From Monday to Wednesday there seems to be an increased activity period during hours 16-20. Similarly Saturdays have increased activity, but during hours 9-14. Sunday does not have an increased activity period, and it is clearly the

shortest day for any activity, which can explain why it has a low level of burned calories.

Summary of Key Findings

The report's findings indicate that total daily activity significantly affects calories burned, with a positive correlation between ‘ActiveMinutes’, ‘TotalDistance’, and ‘Calories’. The analysis also shows that the estimated amount of distance needed to burn a certain level of calories varies, as evidenced by the spread in calorie burn at different distances in the scatter plots.

The activity distribution throughout the week is highest at the beginning and gradually decreases, with the lowest activity on Sundays. This pattern partly aligns with health recommendations for regular activity, but the drop in weekend activity may not be ideal. The findings emphasize the importance of consistent physical activity for calorie burning and overall health.

6. Discussion

Tracking active calories

Trackers provide estimations and should be used as a guide rather than an absolute measure. Understanding the context and limitations of the data is important for effective usage.

Limitations

The precision of calorie tracking devices varies. Wearables and fitness trackers use algorithms and sensors to estimate calories burned, which might not always be precise. Age, gender, weight, and fitness level significantly affect calorie burn. Trackers that take these personal metrics into account tend to be more accurate. Also different activities burn calories at different rates. Trackers may not accurately capture all types of movement, especially non-step-based activities like cycling or swimming.

For this project I believe the major limitation was the consistency in usage. Consistent use of the tracker is essential for accurate long-term data and trends. Unnoticed inconsistent usage can cause major inaccuracies in data-analysis.

Future research

Since there were a significant number of entries with 0 active time and high sedentary minutes, which was very likely caused by not wearing the device at all, I suggest that in future, the entries include a time-based variable that informs whether the tracker has been or not been in use.

7. Conclusion

This project focused on the relationships between different physical activity variables and their health effects. It used Fitbit data to explore patterns of daily physical activity among different population groups. Key findings included a significant positive correlation between total daily activity, calories burned and distance traveled. The study found that activity patterns varied over the week, with higher levels of activity at the beginning of the week and decreasing towards the weekend. The study highlighted the importance of consistent physical activity for calorie burn and overall health. The study also acknowledged the limitations of the tracking equipment and suggested improvements for future research. The project's conclusions highlighted the potential of such data to inform individual health strategies and support global health initiatives to promote physical activity.

References

[1] World Health Organization. (2022). Global Status Report on Physical Activity 2022. Retrieved from https://www.who.int.

[2] Ferguson, T., Olds, T., Curtis, R., Blake, H., Crozier, A. J., Dankiw, K., ... & Maher, C. (2022). Effectiveness of wearable activity trackers to increase physical activity and improve health: a systematic review of systematic reviews and meta-analyses. The Lancet Digital Health, 4(8), e615-e626.

[3] Franssen, W. M., Franssen, G. H., Spaas, J., Solmi, F., & Eijnde, B. O. (2020). Can consumer wearable activity tracker-based interventions improve physical activity and cardiometabolic health in patients with chronic diseases? A systematic review and meta-analysis of randomised controlled trials. International Journal of Behavioral Nutrition and Physical Activity, 17, 1-20.