EXERCISE EXPOSURE DATA PREPARATION FOR CANCER RESEARCH

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

MSc.

by

Ziyu (Andrea) Qiu and Benjamin James Yellin

May 2020

A note about this report: It is the final report for the one-year Specialization Project, required for the Information Systems Masters program with a concentration in Health Technology. This project was a two-person research project done under the advisement of Dr. Deborah Estrin and Dr. JP Pollak at Cornell Tech and Dr. Lee Jones at MSKCC.

ABSTRACT

In this paper, exercise data from 207 cancer patients at Memorial Sloan Kettering Cancer Center (MSKCC) are cleaned and visualized. Previous research has indicated that increased amounts of physical activity can enhance a person's ability to fight cancer, but it is still unknown how much exercise is necessary and how these impacts change depending on tumor type. This project provides a pipeline for data cleaning that can be applied to exercise data from a variety of wearable devices and mobile phones and also suggests ways of visualizing this data. These data visualizations could become useful to clinicians and researchers to identify features of exercise that could potentially be most effective in boosting a person's ability to fight cancer.

CHAPTER 1

INTRODUCTION

In this project, the fitness and sleep data collected as part of the clinical trial Exercise Exposure Data Collection in Patients Undergoing Tumor Molecular Profiling conducted by Dr. Lee Jones' Lab at Memorial Sloan Kettering Cancer Center (MSKCC) are cleaned, organized, and visualized. Previous research has shown that increasing levels of physical activity can impact the rate of breast and colon cancer progression (Clague and Bernstein, 2012). The increasing accessibility and reliability of sleep and physical activity data motivates the research on its use to impact the rate of cancer progression. This includes both how exercise could alter the rate at which tumors progress and how exercise can mitigate the side effects that patients experience from their treatments.

CHAPTER 2

RELATED WORK

Previous studies have found connections between increased levels of physical activity and decreased rates of cancer growth, including exercise's ability to reduce the risk of getting cancer at all (Clague and Bernstein, 2012). For certain types of breast and colon cancers, physical exercise is one of the few factors that individuals have control over that have been shown to reduce the risk of contracting these diseases (Vainio, Bianchini, et al., 2000 in Clague and Bernstein, 2012). Physical activity also has been shown to improve already diagnosed patients? cancer outcomes (Clague and Bernstein, 2012). One research study found that women who exercised moderately after a breast cancer diagnosis had roughly a 40-50% decrease in cancer recurrence as compared with women who exercised less. However, there is still research to be done to determine the optimal amount of exercise a particular person should do in order to minimize the likelihood of their tumor recurring ("Can Exercise Reduce the Risk of Cancer Recurrence?," 2018). Another

study conducted by Jones, Kwan, et al. examined whether certain subtypes of early-stage breast cancer preferentially responded to exercise as a form of treatment (Jones, Kwan, et al., 2016). This analysis was stratified by the clinicopathologic and molecular features of people's tumors. For the unselected cohort, increasing exercise did not reduce the risk of breast cancer recurrence. However, for patients with tumors smaller than 2 cm, patients with well/moderately differentiated tumors, and patients with ER-positive, exercise was found to reduce the number of breast cancer deaths. ER+/PR+/HER2+/low-grade breast tumors responded most readily to exercise as a form of treatment. Work done in both human and mouse models found that voluntary exercise and leisure-time physical activity (LT-PA) decreased the population risk of colon cancer by 13-14%. The notable feature of this study is that this exercise was voluntary and no specific exercise prescriptions were given. Similar results occurred in populations of current and former smokers (Anzuini, et al. 2011 in Clague and Bernstein, 2012). This is of particular relevance to this project because its focus is on people who exercise as a part of their daily life and is not analyzing health outcomes from patients who were prescribed certain levels of exercise. LT-PA was also found to reduce the risk of endometrial cancer in overweight and obese women. Finally, heavy LT-PA was found to reduce the prevalence of various types of lung cancer in former and current smokers. These results show promise for physical activity as a form of protection against cancer, but the specific aspects of exercise which protect against each one remains an area of research (Anzuini, et al. 2011 in Clague and Bernstein, 2013). This project focuses on a particular instance of sleep and exercise data cleaning and visualization and is motivated by previous research showing that improved sleep and frequency of exercise can improve cancer outcomes.

CHAPTER 3

[APPROACH — METHOD — IMPLEMENTATION]

3.1 Observation

207 current or previous patients at Memorial Sloan Kettering Cancer Center granted authorization to researchers to access their exercise and sleep data collected from tracking devices. The data was collected from two sources: health tracking applications and wearable fitness devices (Fitbit, Withings, Misfit Wearables). The data was captured, standardized, delivered to the Validic platform, and exported in JSON format after deidentification.

In order to achieve better organization and visualization, the data was converted from JSON format into one aggregated table. However, 87.6% of data were missing. After observing the data, we found out that the data coming from Validic contained both exercise and sleep data. In order to organize the data, we separated the exercise data and sleep data into different tables based on whether any sleep data was provided for that particular patient, and further subdivided them by specific sources given different levels of granularity.

Activity_alnum contains data from Misfit wearable devices, activity_num contains data from health tracking apps, with detailed information by the breakdown of specific activity type, and activity_date contains data from Withings devices. The data were separated based on the column activity_id, where the format generated by different sources differs based on how each stores data.

Similarly, the sleep data is also separated into subtables based on the column activity_id. Sleep_num provides the richest information, including different stages of sleep

and awake wake cycles during the night, while sleep_alnum and sleep_date only contain the very basic sleep cycle information.

3.2 Data Cleaning

To clean the data, first all columns that had constant null (NaN) values were removed since they provide no meaningful information. Second, extra columns created by Validic (like duplicate timestamps that log when a particular measurement was taken) were discarded. Then if a column had entries that followed the pattern of date and time, its entries were converted into DateTime types.

Some columns that came from Validic were labeled differently but were found to have the exact same values as columns containing variables from people's devices. These repeat data were removed, as they are essentially duplicate features recorded by Validic from those already stored in the fitness applications. Further, columns from Validic that were perfectly correlated with existing data collected from people's devices were noted and discussed with clinicians to determine which units were most relevant. The units were made consistent across different variables. Some of these inconsistencies were due to measurements in minutes versus seconds, and some were due to measurements in meters versus kilometers. Columns with irrelevant units were removed. In collaboration with MSKCC, variables that were relevant to the analysis of physical fitness and sleep were kept and organized into tables. These tables are shown in the Appendix.

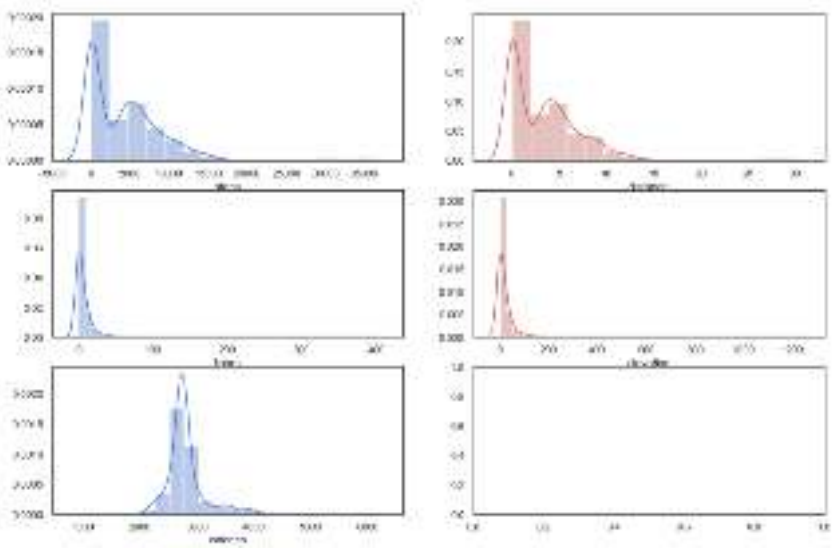

Figure 3.1: These histograms show the frequency of various aspects of walking and running. Both the steps and distance histograms and the floors and elevation histograms look very similar to each other.

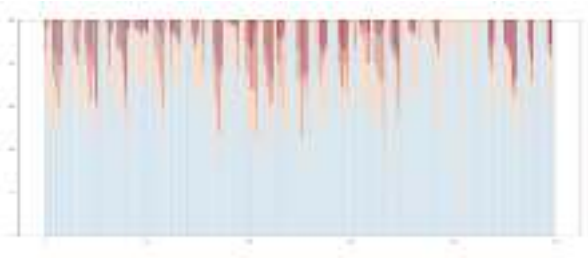

Figure 3.2: This is a heatmap showing the activity level of a single user. Each single vertical line corresponds to the activity data available for a single day. Blue indicates a low level of activity, orange indicates a medium level of activity, and red indicates a high level of activity.

3.3 Visualizations

Visualizations were done on a Fitbit dataset available online as well as on the exercise and sleep data from MSKCC. The analyses done on both the population and individual level are shown in Figures 3.1 through 3.10.

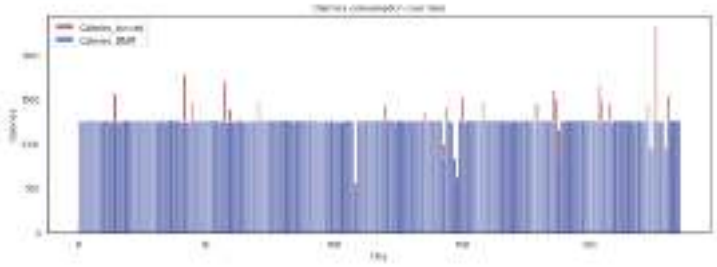

Figure 3.3: The blue in this plot indicates the basic calories burned without doing any intentional physical activity, and the red indicates active calories burned. From this graph, we can observe both the total calories burned daily, and how many calories are burned by activity.

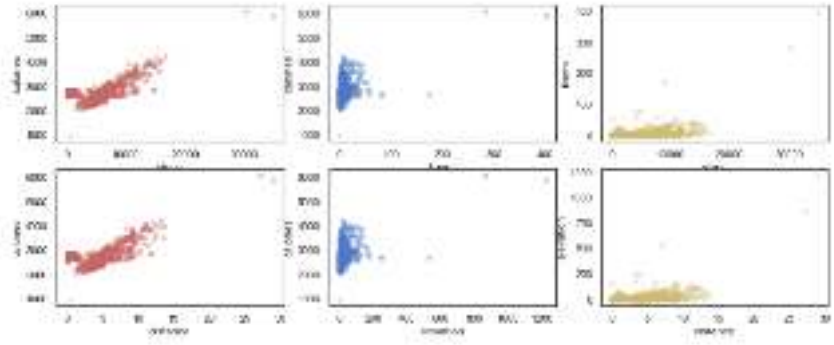

Figure 3.4: These graphs give an overview of how the number of calories burned varies with other exercise features that were recorded.

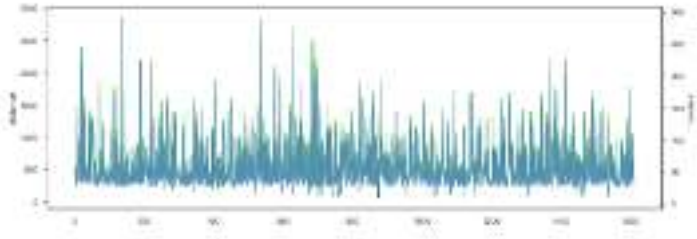

Figure 3.5: This plot is for a single patient, where each line shows the number of calories burned (blue) and the distance traveled (green) captured by one data entry. These distributions generally follow the same distribution pattern.

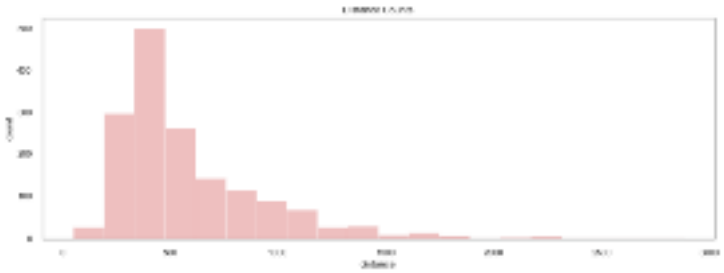

Figure 3.6: This distribution shows how many times a single patient traveled various distances.

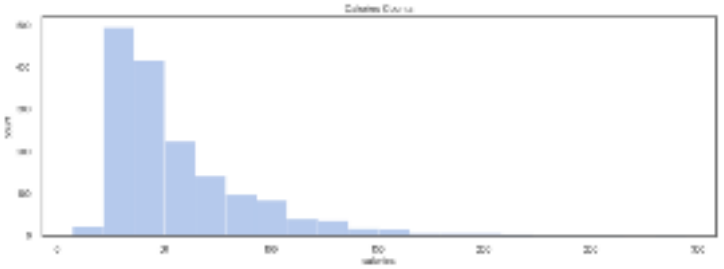

Figure 3.7: This distribution shows how many times a single patient burned various amounts of calories.

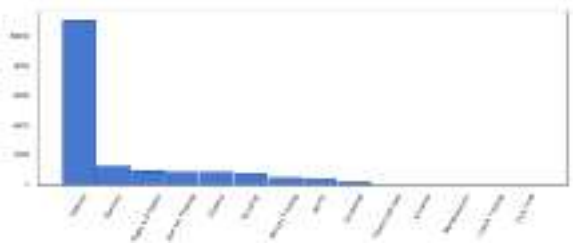

Figure 3.8: This distribution shows how many times a single patient participates in various different types of exercise.

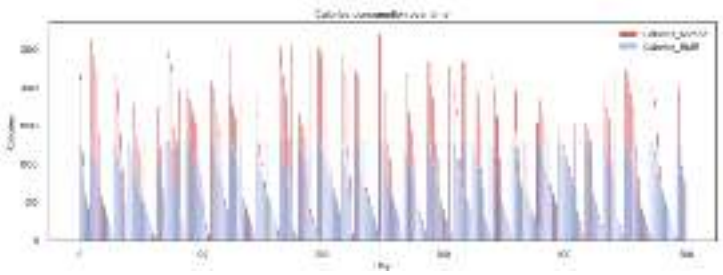

Figure 3.9: This plot shows a single patient's calories burned and basal metabolic rate over several days.

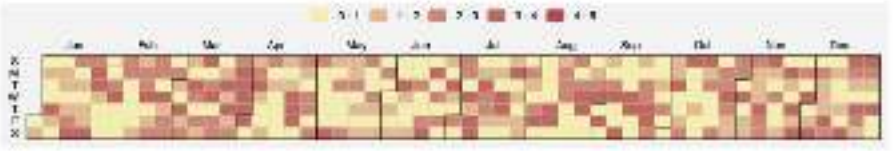

Figure 3.10: This is a heatmap of a user's sleeping status shown on a calendar. The color indicates the number of disruptions experienced during the night. Yellow refers to lack of disruptions, and darker shades of red refer to more disruptions.

CHAPTER 4

RESULTS

The output of this project is organized and visualized sleep and exercise data from 207 cancer patients. Before cleaning and visualizing data from cancer patients at MSKCC, visualizations were also generated for Fitbit data that is freely available online. Then, the techniques applied on the freely available dataset were translated to the data from MSKCC in order to see how frequently patients participated in various levels of physical activity and how frequently their sleep was of various qualities.

CHAPTER 5

DISCUSSION

The visualizations performed on exercise and sleep data that are shown here can eventually be used to identify meaningful features of each of these activities that can be incorporated into a model that predicts the rate of cancer progression. In order to begin to identify these meaningful features, it is necessary to understand how various features of exercise are related to each other so that two variables encoding the same information would not be used in the same model. For this reason, visualizations are generated that demonstrate how certain variables are related to each other. This is done either by plotting two variables against one another or by examining the distributions of the variables side by side.

Sleep data like what is shown in Figure 3.10 could be analyzed along with disease progression data in order to determine whether patients who are not able to sleep as well are having worse treatment outcomes. It is also possible, however, that patients would not be able to sleep as well if their cancer was progressing more rapidly, so a causality cannot be directly inferred from the correlation. If biomarkers of a person's disease were taken periodically, it could also be analyzed whether people were more responsive to a particular cancer treatment if they had slept well recently. If patients consistently slept well on particular days, this could be used to schedule cancer treatments.

CHAPTER 6

CONCLUSION

This project focused on building a pipeline for cleaning and visualizing sleep and exercise data from cancer patients in order to help clinicians and researchers better understand the pattern of exercising and sleeping of their patients on both population level and individual level. This will eventually be used to build a model of cancer progression that incorporates certain aspects of exercise and sleep quality as parameters.

6.1 Limitations

The procedure followed in this project for cleaning and visualizing Fitbit data could be applied in other contexts, but in different populations, the visualizations themselves would look different. The use of descriptive statistics to capture all of a single patient's data in one row of a table helps to significantly reduce the dimensionality of the data, but also decreases the granularity of information available about that patient, which would likely be necessary to predict features of their cancer. Future work would focus on feature selection to determine how best to describe a patient's exercise and sleep data. Without labelled data about disease progression, patient survival, or some tumor metrics, it is not possible to make specific conclusions about the extent to which an increase in exercise is correlated with a slowing of cancer progression.

6.2 Future Work

The pipeline developed in this project can be used to transform data from fitness devices and applications into a format that is more easily usable for research purposes. Future work should utilize the visualizations performed in this project and focus on identifying meaningful features relating to sleep and exercise quality and quantity that can be used to predict the rate of cancer progression. One aspect of this project was the visualization of time series data for a single patient. Another avenue for further exploration is clustering of patients based on specific features of importance. This could provide a way of stratifying the population and enable researchers to explore these subpopulations further.

REFERENCES

Clague, Jessica and Bernstein, Leslie. "Physical Activity and Cancer." *Current Oncology Reports*(2012): 550-558.

Vainio, H. "Weight Control and Physical Activity." *IARC Handbooks of Cancer Prevention*,*6* (2002).

Jones, L. "Precision Oncology Framework for Investigation of Exercise As Treatment for Cancer." *Journal of Clinical Oncology*,33, no.35, Dec. 2015, pp. 4134-4137. http://ascopubs.org/doi/10.1200/JCO.2015.62.7687

"Can Exercise Reduce the Risk of Cancer Recurrence?" *Dana-Farber Cancer Institute*,7 Feb.
2018, blog.dana-farber.org/insight/2018/02/can-exercise-reduce-risk-cancer-recurrence/.

Anzuini, F., Battistella, A., Izzotti, A. "Physical activity and cancer prevention: a review of current evidence and biological mechanisms." *Journal of Preventative Medicine and Hygiene*, 52, no. 4, Dec. 2011, pp. 174-180. https://www.ncbi.nlm.nih.gov/pubmed/22442921.

Jones, Kwan, et al. "Exercise and Prognosis on the Basis of Clinicopathologic and Molecular Features in Early Stage Breast Cancer: The LACE and Pathways Studies." *Cancer Research*,76, no. 18, Aug. 2016, pp. 5415-5422. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5026589/.

Iacopetta B. Are there two sides to colorectal cancer? Int J Cancer. 2002;101(5). in Clague and Bernstein, 2013, pp.403-408. https://www.ncbi.nlm.nih.gov/pubmed/12216066.

Morikawa, T. et al. "Association of CTNNB1 (beta-catenin) alterations, body mass index, and physical activity with survival in patients with colorectal cancer" *JAMA*vol. 305,16 (2011): 1685-94.

Uhm, Kyeong E., et al. "Effects of Exercise Intervention in Breast Cancer Patients: Is Mobile Health (mHealth) with Pedometer More Effective than Conventional Program using Brochure?"*Breast Cancer Research and Treatment*, vol. 161, no. 3, 2017, pp. 443-452. *ProQuest*, https://search.proquest.com/docview/1859049185?accountid=10267,

doi:http://dx.doi.org/10.1007/s10549-016-4065-8.

Dela Cruz, M.A., Roy, P., Chowdhury, S., Chan, S., Roy, H.K. "Exercise and triple negative

breast cancer: Unravelling the anti-neoplastic molecular factors through novel culture method."
Cancer Research.Feb. 2017,
http://cancerres.aacrjournals.org/content/77/4_Supplement/P3-07-18.

Idorn, M., thorn Straten, P. "Exercise and cancer: From 'healthy' to 'therapeutic'?" *Cancer Immunology Immunotherapy* vol. 66, no. 5, 21 March 2017, pp. 667-671. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5406418/.