Paper RW02

The promise of wearable technology in clinical trials: a proof of concept to deliver it?

Jennifer Bradford, PHASTAR, Macclesfield, United Kingdom
Jonathan Murray, PHASTAR, Glasgow, United Kingdom

ABSTRACT

Consumer grade wearable devices offer the potential to monitor many different physiological measures of health and fitness continuously as individuals go about their daily routines. For a clinical trial this could provide valuable insights between hospital visits, potentially enhancing the understanding of treatment response, delivering a rudimentary early warning system, providing objective measures of more subjective outcomes or even provide efficiencies in trial conduct.

Before the widespread adoption of such technology on clinical trials, however, there are several significant challenges from data access and interpretation to clinical validation. In this paper we describe a study using FitBit Inspire HR devices and offer insights into our experience of some of the specific challenges that exist focused on the collection and analysis of data including the interpretation and generation of meaningful clinical insights.

INTRODUCTION

There is considerable discussion around the use of patient-generated data in health research and clinical studies, particularly from wearable activity trackers. Since around 2013 there has been a large increase in the sales of consumer grade activity tracking devices and it is estimated over 22% of adults in the US alone will wear a smart watch for at least 1 month in 2019 according to eMarketer and the smart wearables market is predicted by Forbes to double by 2022 (October 2018). This popularity gives rise to a potentially underutilised data source for clinical research.

During a clinical study a patient is often closely monitored with a detailed analysis of their blood, vital signs etc. at specified timepoints or visits to the clinic. Basic activity trackers provide accelerometer-based activity data, with more sophisticated models also capable of monitoring heart rate and other features. Such measures have the potential to provide insights into an individual's health, fitness and sleep 24 hours a day, 7 days a week and on a clinical trial this would provide valuable insights for each patient between clinic visits. Furthermore, there are subjective events which can be both under reported and difficult to measure, for example fatigue or general well-being. From a physician's perspective, a greater awareness of these events may help them with appropriate interventions or advice and more generally this information could be important in defining the tolerability of new treatments from a patient's perspective. Wearables offer the potential for objective measures of more subjective outcomes such as fatigue and/or novel endpoints.

Despite the hype around wearables there is limited evidence of how the data from consumer grade devices might be used in clinical studies. Although wearable activity trackers are gaining traction in clinical trials, most focus on their use either as an intervention (e.g. in weight loss or behavioral studies) or in comparing consumer grade with medical grade devices. There is potential for clinical teams to extract value from this data alongside the routinely collected data at visits. It is worth noting that there is much debate around the use of consumer grade devices in clinical research and the accuracy and validity of most devices has not been verified formally. As a result, their use within such research should be with the appropriate level of caution/awareness.

Consumer grade activity trackers may have the potential to provide insights into patients' general health and fitness, and overall well-being between visits. There are, however, several challenges associated with the access to and utilisation of this type of data for instance:

1. Data connectivity: accessing and extracting the data for an individual
2. Types and frequency of data reported
3. Missing Data (including non-wear time)
4. Generating meaningful insights from the data
5. Data privacy and security
6. Clinical validation

Despite these challenges consumer grade wearable data may provide valuable information alongside the standard clinical data collected during a clinical trial. Moreover, it is recognised that actively involving patients in clinical research can improve recruitment and raise the quality of the findings. Empowering patients to donate their own data toward decision making about their own health may have a positive impact on their own experience of a clinical study and may improve the dialog between patient and caregiver.

In this paper we discuss the first 4 of the challenges above; effective extraction, processing and creating actionable health-related insights, and we suggest possible future steps. Overcoming these challenges will enable continuous monitoring of individuals outside of the clinic potentially. This could empower the patient as they will retain control over the additional information and data they provide to the clinical study team.

METHODS

STUDY DESIGN

A total of 15 subjects were recruited, all UK based PHASTAR employees who volunteered to participate. Each volunteer was provided with a FitBit Inspire HR device for the duration of the study. They were asked to wear the device on their non-dominant arm for a total of 1 month during July 2019. Each volunteer created their own private FitBit account and only the individual had access to that account.

In addition, volunteers were asked to complete an online version of the World Health Organisation's five well-being index (WHO-5) [1, 2] which is a short self-reported measure of current mental well-being over the previous 2 week period. This was created in Microsoft Office 365 forms. During the 1-month period, participants were asked to complete the WHO-5 questionnaire 3 times: at the start, after 2 weeks and at the end.

DATA COLLECTION

On completion of the study the participants downloaded their summarised daily activity sleep and activity data from the FitBit website and saved it anonymously on the PHASTAR internal shared drive.

The WHO-5 questionnaire data was exported to Microsoft Excel from Microsoft Office 365 forms and the total score calculated as described below. All data was stored on the internal PHASTAR shared drive with restricted access and each dataset referenced by a volunteer ID rather than their name.

DATA ANALYSIS

All analyses were performed in RStudio [3] or SAS®. Following manual data extraction by each volunteer and subsequent data upload onto the internal shared drive the activity, sleep and survey data were analysed.

SLEEP START AND END

Given that the sleep and activity are reported separately we must decide whether the 'date' of the sleep activity is the period that precedes or follows the activity. For instance, for a given activity completed during Monday, should the sleep precede or follow the activity (e.g. Sunday evening through to Monday morning or Monday night through to Tuesday morning respectively) In our analysis we considered sleep quality following activity on a given day, hence we chose that sleep would follow activity.

PROCESSING THE WHO-5 QUESTIONNAIRE RESULTS

The WHO-5 questionnaire consists of five statements, which respondents rate according to the following scale in relation to the past two weeks.

- All of the time = 5
- Most of the time = 4
- More than half the time = 3

- Less than half the time = 2
- Some of the time = 1
- At no time = 0

The five statements are as follows:

1. Over the last 2 weeks I have felt cheerful
2. Over the last 2 weeks I have felt calm and relaxed
3. Over the last 2 weeks I have felt active and vigorous
4. Over the last 2 weeks I woke up feeling fresh and rested
5. Over the last 2 weeks my daily life has been filled with things that interest me

The total raw score ranges from 0 to 25 and is multiplied by 4 to give the final score. 0 represents the worst imaginable well-being and 100 represents the best imaginable well-being [4]. For each volunteer at each of the three timepoints the final score was calculated.

NON-WEAR TIME CALCULATION

For a given date non-wear time was calculated as shown below using the total minutes in a day (1440), the number of active minutes, the number of minutes asleep at the start of the day (midnight to wake up time) and time asleep at the end of the day (bed time to midnight, if after midnight this time was taken from the following day calculation).

$$non\ wear\ time = \ total\ minutes\ in\ a\ day - (active\ minutes + midnight\ to\ wake\ up\ time + bed\ time\ to\ midnight)$$

Where

$$active\ minutes = sedentary\ minutes + lightly\ active\ minutes + fairly\ active\ minutes + very\ active\ minutes$$

An example is given below as an example for the 2nd July:

Information reported by FitBit:

- *Sleep time 10:00 pm on 1st July to 7:00 am on 2nd July*
- *Sleep time 10:00 pm on 2nd July to 7:00 am on 3rd July*
- *Sedentary minutes on 2nd July = 500*
- *Lightly active minutes on 2nd July = 270*
- *Fairly active minutes on 2nd July = 40*
- *Very active minutes on 2nd July = 30*

Calculation

- *Total active minutes on 2nd July = (500 + 270 + 40 + 30) = 840 minutes*
- *Minutes asleep at the start of day = minutes between midnight and 7am on 2nd July = 420 minutes*
- *Minutes asleep at the end of day = minutes between going to bed and midnight on 2nd July = 120 minutes*

- *Total wear time for 2nd July = (840 + 420 + 120) = 1380 minutes*
- *Total non-wear time = total minutes in 1 day – total wear time = 1440 – 1380 = 60 minutes*

SLEEP QUALITY

The method of Valenti et al. [5] was used to calculate sleep quality. It was calculated for each sleep period using the FitBit reported sleep stages as

$$sleep\ quality = \frac{minutes\ in\ REM\ sleep + minutes\ in\ deep\ sleep}{total\ number\ of\ minutes\ asleep}$$

MODELLING THE WHO-5 SCORE

A Multiple regression with repeated measures model using SAS® MIXED procedure was used to fit WHO-5 score against the FitBit variables. A stepwise selection method was used with effects entering or exiting the model based on the significance level primarily.

The WHO-5 score was adjusted for the baseline score by calculating the change from baseline. The FitBit data was averaged over the previous 14 days meaning each volunteer contributed 2 observations to the modelling dataset. As sleep variables were included as possibilities in the fitting process, the SAS® MIXED procedure excluded one of the 30 observations due to missing sleep data.

The AIC (Akaike Information Criterion) score [6] can be used to estimate the relative quality of models, and was further refined to the AICc score to correct for small sample sizes [7, 8]. AICc was used to assess the relative quality of the models produced.

RESULTS

As discussed in the introduction there are many challenges in utilising the data from consumer grade activity trackers, exacerbated if those devices belong to an individual rather than being supplied as part of the trial:

- Data connectivity: accessing and extracting the data for an individual.
- Types and frequency of data reported:
- Missing Data (including non-wear time)
- Generating meaningful insights from the data
- Data privacy and security
- Clinical validation

In the following sections we discuss each of the first 4 challenges and describe our experience using the FitBit Inspire HR devices.

CHALLENGE 1: DATA CONNECTIVITY

One of the biggest challenges is accessing the data from these devices. There are two different challenges here, the first is what the manufacturer will allow you to access, and the second is how to create a workflow for secure transfer of the data from the device for analysis.

In this study we provided full instructions so that the volunteers could download their data to a secure server. They were able to select what data to download and all participants selected to share all of their sleep and activity data. This approach is not feasible on a larger study of course. FitBit provided data at the daily summary level through this mechanism. It is possible to access the intraday data through the creation of an Application Programming Interface (API), however this would be subject to review by the device company.

The approach taken here was a simple approach, relying on the volunteers to share their data with the Investigators. Scaling to a larger clinical trial would require careful consideration around how, with appropriate consent, the data for an individual would be securely transferred, anonymised, and linked to existing data. Also, consideration must be given to the appropriate laws governing different parts of the world, for instance in the EU it would be GDPR, whereas in the US there are different laws depending on the grade of the device.

CHALLENGE 2: TYPES OF DATA REPORTED

As described previously, although activity trackers essentially measure activity via an accelerometer and possibly a heart rate monitor there are differences in the data available from different devices and any derivations from the device company. Below is a summary of all the data/measurement types that was provided in the data download from the FitBit dashboard for the FitBit Inspire HR. Each measurement was a daily summary:

<u>Activity</u>

- Calories burned
- Steps
- Distance

- Minutes sedentary
- Minutes lightly active
- Minutes fairly active
- Minutes very active
- Activity calories

Based on Sleep

- Start and end time of sleep
- Minutes asleep
- Minutes awake
- Number of awakenings
- Time in bed
- Minutes REM sleep
- Minutes light sleep
- Minutes deep sleep

Every volunteer provided this information as a CSV file. Limitations to using this data include: the data is summarised as daily values, the definition of activity and sleep rely on FitBit's interpretation of the heart rate data, and we are not provided with any accelerometer or heart rate data. These device specific limitations must be considered when we have chosen the data we wish to analyse.

CHALLENGE 3: MISSING DATA AND NON-WEAR TIME

There are 2 types of missing data with the activity trackers. The first is where some data for a given day is reported but the granular data is unavailable. An example is sleep data: the time to bed and wake up times maybe reported but the sleep stages may be missing. Alternatively, we may have activity data for each day but no sleep data at all, suggesting an individual removed their device at night. Figure 1 shows the amount of missing data across the different volunteers for each category of data. This suggests that most of the missing data was a result of the individuals removing the device at night (to sleep).

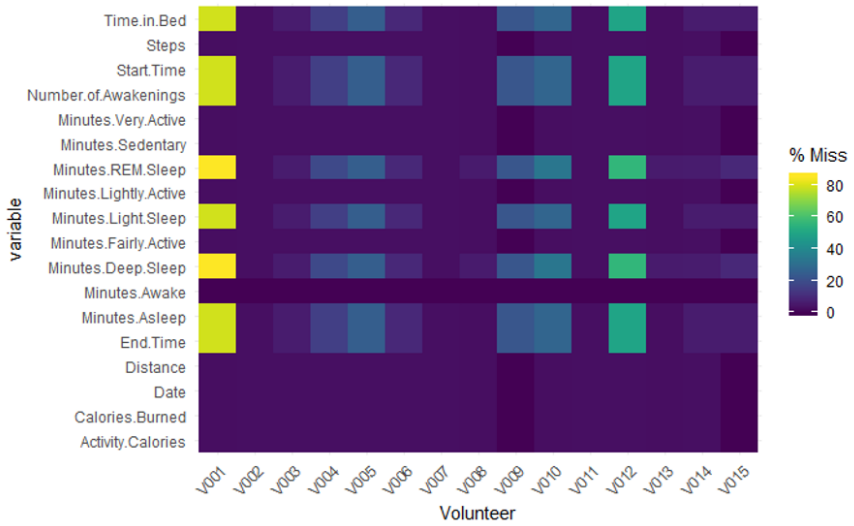

Figure 1 The percentage of missing data for the different categories of data reported from the FitBit Inspire HR across the volunteers on the study for the duration of the 1 month study.

The second challenge around missing data is non-wear time, this is less obvious. FitBit does not directly report non-wear time, however from the data it was possible to approximate non-wear time per day for an individual based on a

calculation over a 24-hour period. The calculation is described in the methods section and was based on two assumptions:

(1) an individual does not wake between the sleep start and end time (this of course is not strictly correct as individuals could wake in the night and remove the device for a time and then put it back on).
(2) If FitBit reports the number of sedentary minutes for a given day to be 1440 (the total number of minutes in a day) then it was assumed the user was not wearing the device

In this study we looked at the non-wear time over the month. In general, as shown in Figure 2, the total non-wear time was low. Questionnaire results suggest most individuals only removed the device for charging and showering, with some choosing to not wear the device for prolonged periods such as for particular social occasions or to sleep.

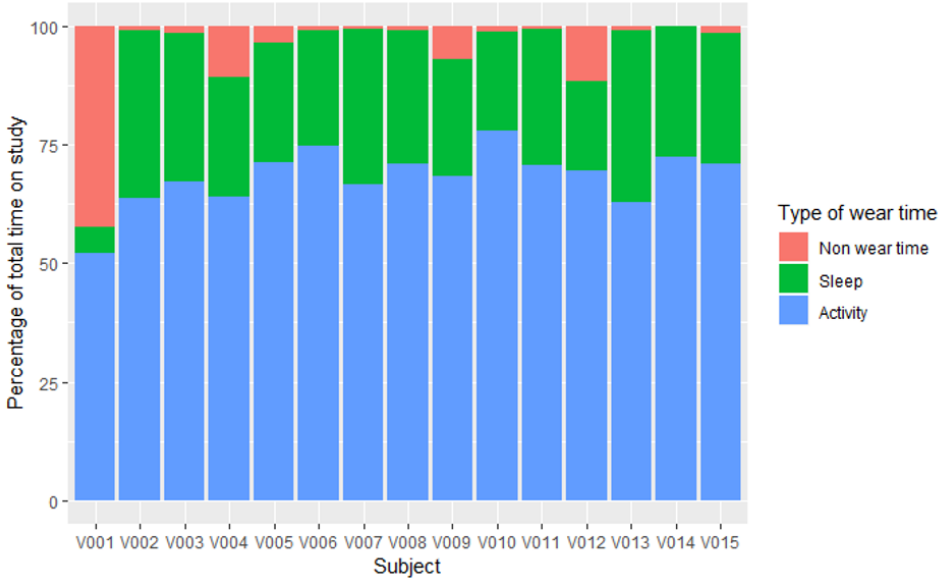

Figure 2 For each subject for the duration of the 1 month study the percentage of sleep, active minutes and approximate non wear time is plotted

Any further analyses should factor in non-wear time and missing data. It is not yet determined how representative this is over the broader population. When considering this information on a larger study thought should be given to what missing data there might be, how it will be calculated or detected and how to deal with the missing data. Certainly, the amount of missing data should be made clear to anyone interpreting the data along with any assumptions that have been made.

CHALLENGE 4: GENERATING MEANINGFUL INSIGHTS

INSIGHTS DIRECTLY FROM THE FITBIT DATA

One major challenge in the use of these devices is around the interpretation of the data by clinicians. The volume of data generated by wearables could potentially be overwhelming to a physician or a clinical study team. For instance, many trackers can measure heart rate continuously over a 24 hour period, which over the course of a clinical study is a lot of data points. There is still much to be done in this area in understanding how this data can be translated into something usable by the clinical and/or study team. Moreover, it is important to consider what medical intervention would result from the interpretation of such information.

It is known that sleep is an important measure of health [9], activity can improve aspects of our well-being [10] and that physical activity impacts the overall quality of sleep [11]. If we look at the data generated at the single patient level it may be possible to look at changes in the sleep of an individual over time or a decrease in activity levels. The result from such insights may lead to an enhanced conversation between patient and caregiver with two fold impact; firstly the patient may feel more reassured that they are contributing to the consultation and secondly, there may be a medical

intervention or a specific piece of advice the caregiver can provide that may help that patient.

Figure 3 shows an example of the total minutes asleep over the duration of the study for 4 volunteers on the study. Although each subject is different, with some having a more regular sleep pattern it is difficult to imagine how a physician would interpret the results. The volunteers in this study were all healthy and there were no significant changes over the course of the study. For some individuals it may be possible to detect a general change overtime by sight and relate this to other symptoms being reported. Figure 3 also demonstrates how varying amounts of missing sleep data can impact the results and make it increasingly more difficult to extract anything useful.

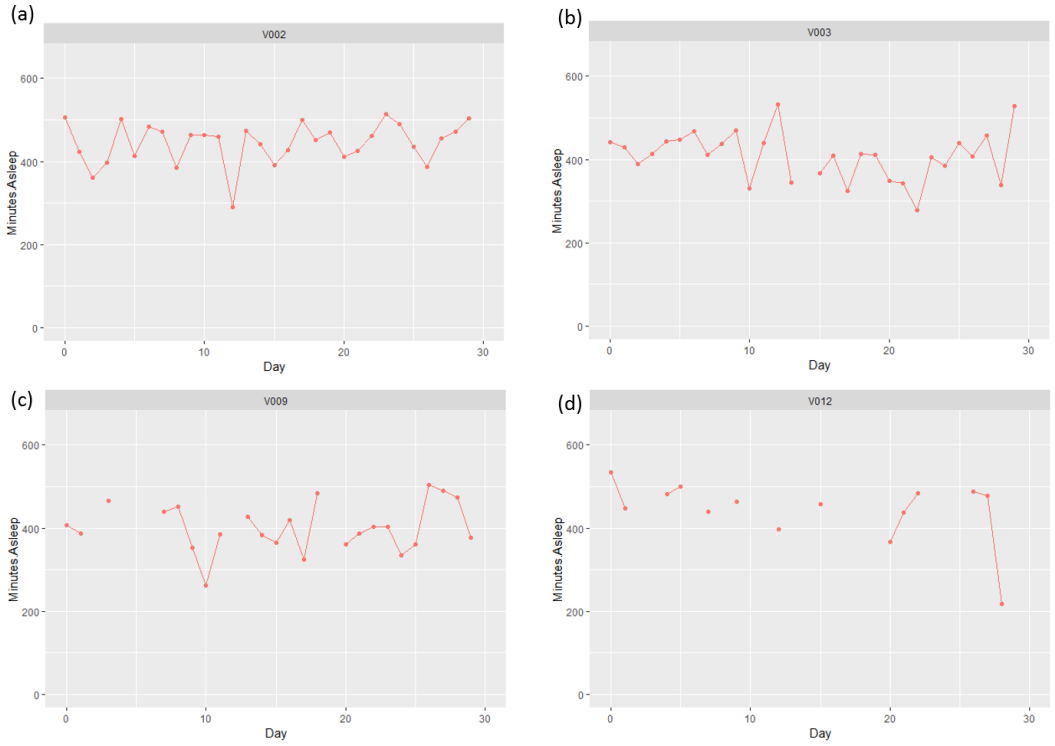

Figure 3 The minutes asleep as reported by FitBit for 4 subjects from the study over the course of the study.

A clinical study team may be interested in the data at the level of a population and want to consider potential trends in the data to provide meaningful insights. There are many ways to look at the data and in Figure 4 we look at the non-sedentary time on each day of the week. The core hours of the company are Monday to Friday and as expected there is a little more activity at weekends. Translated to a clinical study this sort of information could look at trends following dosing, for example the general activity levels post-dose across the trial population, which may provide insights into tolerability.

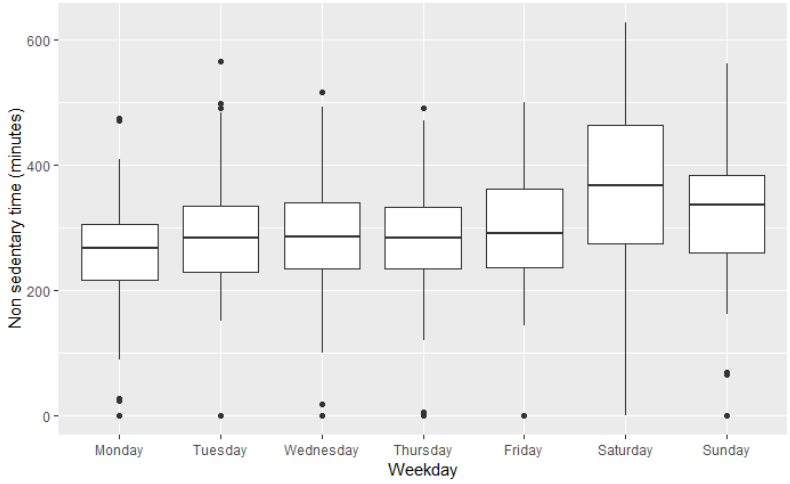

Figure 4 The non-sedentary time (lightly active + fairly active + very active as described by FitBit) across the 15 subjects on the study for each day of the week.

More generally we can look at the association between activity and sleep. There is no consensus on how to measure sleep quality. Here we apply the approach of Valenti et al. [5] and define sleep quality as the time spent in REM and deep sleep divided by the total minutes asleep and look at activity vs sleep quality. Figure 5(a) shows the non-sedentary time in minutes against the sleep quality and Figure 5(b) the very active minutes against the sleep quality; there is no correlation evident in either relationship.

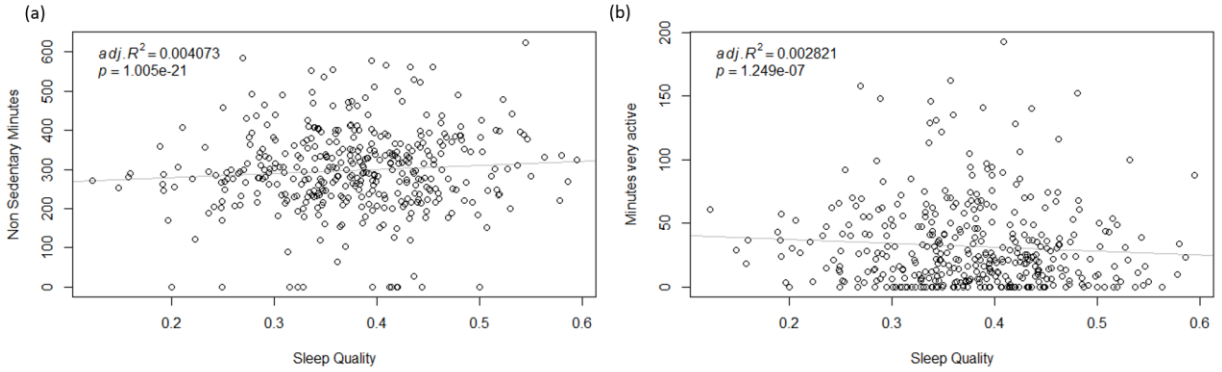

Figure 5 The sleep quality vs. (a) the daily number of non-sedentary minutes (lightly, fairly and very active minutes) and (b) the daily very active minutes as reported by FitBit.

Clearly, the volume of data from activity trackers is potentially very large and it isn't clear how a caregiver may extract value from such data in a consistent way for an individual subject. It is worth noting here that even in this study where only daily summary data was used there were still many data points created during the study. It may be more beneficial to use the activity/sleep data as an indicator of other health events. One potential application could be in the indication of a general sense of a patient's well-being, which in turn could provide insight into drug tolerability.

Well-being is subjective, sometimes relying upon a dialog between patient and caregiver or a questionnaire completed by the patient. One approach used to measure well-being is a validated questionnaire, completed by the patient which relies upon recall from a previous fixed duration. One such questionnaire is the WHO-5 [1, 2], which is a short self-reported measure of current mental well-being over the previous 2 week period. During the study we requested participants to complete an online version of the WHO-5. Although not a large enough sample size to be statistically meaningful it provides some preliminary data on any possible association of activity tracking data to well-being.

FITBIT DATA AS A POTENTIAL OBJECTIVE MEASURE OF WELL-BEING

Multiple regression with repeated measures using SAS® mixed procedure was applied to the activity and sleep data and WHO-5 scores. The first model (Model A) initially included all variables and effects with highest p-value (least significance) were removed systematically (see Table 1). Collinearity (where some of the independent variables are highly correlated) can indicate that the estimates for the coefficients are unstable; collinearity should be limited where possible. Model A includes 3 variables with high correlation (mean number of minutes asleep, mean number of minutes in deep sleep and mean percentage of deep sleep). To address this, a second model (Model B) was adapted from the first model with the goal of replacing the 'sleep' variables with a single variable. The second model included: baseline WHO-5 score, mean number of calories, mean time in bed (minutes), and median steps (see Table 2).

Effect	Estimate of Coefficient	Standardised Estimate of Coefficient	P-value (5% Significance level)
Intercept	379.50	-0.01981	0.0029
Baseline WHO-5	-0.7464	-0.5725	0.0007
Mean Calories Burned	-0.01594	-0.4286	0.0062
Mean Minutes Asleep	-0.9003	-2.8852	0.0072
Mean Minutes Deep Sleep	4.8838	3.6535	0.0201
Mean Percentage Deep Sleep	-17.8935	-2.8826	0.0244
Median Steps	0.003917	0.6561	0.0004

Table 1 Estimates for coefficients for Model A. (AICc score for Model A was 228.5)

Effect	Estimate of Coefficient	Standardised Estimate of Coefficient	P-value (5% Significance level)
Intercept	108.53	-0.02602	0.0022
Baseline WHO-5	-0.7498	-0.5751	0.0021
Mean Calories Burned	-0.01343	-0.3611	0.0410
Mean Minutes Time in Bed	-0.1331	-0.5054	0.0056
Median Steps	0.003152	0.5278	0.0031

Table 2 Estimates for coefficients for Model B. (AICc score for Model B was 238.9)

The AICc score allows us to compare how well the model fits the data (smaller is better). Whilst Model A fits the data better than Model B, it may not be a better model for prediction; given the low number of observations Model A may be over-parameterised in comparison to Model 2. A simpler model with fewer terms is usually preferred. By substituting the three 'sleep' variables with one in Model B, we have lowered the collinearity in our model variables. The fit of Model B to the data may be worse (AICC 228.5 -> 238.9), but we have a simpler model that may provide more flexibility when predicting WHO-5 scores for additional subjects.

In addition to a potential model providing an objective measure of well-being such a model may also provide valuable insights into how the different variables affect the WHO-5 score. The coefficients of the models describe how the outcome variable (adjusted-for-baseline WHO-5 score (WHO-5b)) will be affected when we increase the explanatory variable by one unit while keeping all other variables constant. The standardised estimates allow us to more easily see the relative impact of the variables included in the model; the effect whose standardised coefficient has the largest absolute value is most impactful.

Interpreting the coefficients in Model A is challenging due to the high collinearity of the 'sleep' variables; changing just one variable without changing another would be unlikely. In Model B (Table 2) the standardised estimates tell us that a subject's WHO-5b score would be most impacted by increasing their number of daily number of steps followed by reducing their time in bed. The estimate column can be interpreted such that increasing median daily steps by 1000 will raise the WHO-5b score by 3.1 points; an expected observation given that common sense tells us that increasing activity has an impact on our well-being. However, the model also suggests that reducing the mean time in bed by 30 minutes will raise the adjusted-for-baseline WHO-5 score by 4 points approximately, which is unexpected as one might assume that a longer time in bed would be beneficial.

We have described how it could be possible to generate meaningful, objective measures from consumer grade activity trackers, this approach would benefit from a larger dataset where it may be possible to create a model that could be

validated against a measure such as a WHO-5 score to provide an objective measure of well-being. Such a model would also provide valuable insights into the variables that effect the model which may enhance our understanding of well-being more generally.

CONCLUSION

Consumer grade wearable devices offer the potential to monitor an individual 24/7 as part of their daily routines, away from the hospital and include a myriad of physiological measurements from brain activity through to temperature. Although there are several clinical trials utilising wearable devices, often as an endpoint or as an intervention, researchers are only just beginning to understand how to extract value from such rich data to generate clinical insights. For instance, digitally collected data from a wearable device may be transformed, through mathematical models and machine learning, into indicators of health outcomes. It may be possible that characteristics in an individual's activity or sleep could form an objective measure of their general well-being or fatigue levels which could be used to inform their physician and/or provide insights across the study around general tolerability.

The role of wearable technology in healthcare and clinical trials specifically is promising; the potential applications are diverse including enhancing our understanding of responses to treatment and efficiency gains in clinical trial conduct. However, there are significant challenges and the community would benefit from the regular exchange of learnings and experiences to facilitate the development of best practices around the deployment, collection, standardisation and analysis of such data for validated clinical applications.

REFERENCES

[1] C. W. Topp, S. D. Østergaard, S. Søndergaard and P. Bech, "The WHO-5 Well-Being Index: A Systematic Review of the Literature.," *Psychotherapy and Psychosomatics,* vol. 84, pp. 167-176, 2015.

[2] WHO, "Wellbeing Measures In Primary Health Care/The Depcare Project," WHO Regional Office for Europe, Copenhagen, 1998.

[3] RStudio team, "Rstudio: Integrated Development for R," http://www.rtudio.com, Boston, MA, 2015.

[4] Clinical Outcomes Research Consortium (CORC), "The World Health Organisation- Five Well-Being Index (WHO-5)," [Online]. Available: https://www.corc.uk.net/outcome-experience-measures/the-world-health-organisation-five-well-being-index-who-5/.

[5] G. Valenti, A. G. Bonomi and K. R. Westerterp, "Quality Sleep Is Associated With Overnight Metabolic," *J Gerontol A Biol Sci Med Sci,* vol. 72, no. 4, pp. 567-571, 2017.

[6] B. N. Petrov and F. Csaki, "Information Theory as an extension of the maximum liklihood principle," in *Second International Symposium on Information Theory*, Budapest, 1973.

[7] K. Burnham and D. Anderson, Model Selection and Multimodal Inference, New York: Springer, 2002.

[8] C. Hurvich and C. L. Tsai, "Regression and time series model selection in small samples," *Biometrica,* vol. 76, pp. 297-293, 1989.

[9] R. Cooke, "Sleep should be prescribed': what those late nights out could be costing you," 24 September 2017. [Online]. Available: https://www.theguardian.com/lifeandstyle/2017/sep/24/why-lack-of-sleep-health-worst-enemy-matthew-walker-why-we-sleep. [Accessed September 2019].

[10] "How to look after your mental health using exercise," [Online]. Available: https://www.mentalhealth.org.uk/publications/how-to-using-exercise.

[11] P. D. Loprinzi and B. J. Cardinal, "Association between objectively-measured physical activity and sleep, NHANES 2005–2006," *Mental Health and Physical Activity,* vol. 4, no. 2, pp. 65-69, 2011.

ACKNOWLEDGMENTS

The authors would like to thank the volunteers for participating in the study along with Susan Lovick, Jennifer Rogers and Alastair Sword from PHASTAR who provided invaluable advice.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Jennifer Bradford
PHASTAR
Unit 2A 2 Bollo Lane
London W4 5LE
Email: Jennifer.bradford@phastar.com
Web: https://www.phastar.com