STATISTICAL ANALYSIS PLAN

USING ACTIVITY TRACKING AND JUST-IN-TIME MESSAGING TO IMPROVE ADAPTIVE PACING: A PRAGMATIC RANDOMISED CONTROL TRIAL

Statistical Analysis Plan
(SAP)

Version: 1

Author: Dr Lawrence Hayes, Lancaster University Medical School

Date: 16/07/2021

Contents

1. INTRODUCTION

People with long COVID report 'push-crash' cycles, with a disproportionate worsening of symptoms in response to activity, similar to post-exertional malaise (PEM) reported by patients with chronic fatigue syndrome [1–3]. These PEM-like symptoms increase the overall symptom load, reduce each individual's quality of life, and make recovery harder.

Adaptive pacing has emerged as a common strategy to self-manage PEM-like symptoms [4,5]. However, recent work endorsed by our long COVID PPI group indicates that implementing adaptive pacing is often problematic for those attempting to self-manage. Difficulties include accurately estimating concepts such as energy availability and predicted energy use and identifying and tracking a suitable threshold to limit activity. Combining these requirements to plan daily activities, often hour by- hour, can be highly challenging, particularly when symptoms include impaired cognition.

This project aims to determine if combining continuous activity tracking with a just-in-time adaptive intervention (JITAI) can address these limitations. JITAIs provide information to participants at a time and context where they can act upon it [6,7]. Widely used in behaviour change research, JITAIs are yet to be applied to adaptive pacing. Using a randomised control study design, we will allocate 250 participants to receive either JITAI supported adaptive pacing or usual care. Our primary outcome is PEM using the De Paul Symptom Questionnaire (DSQ)-PEM [8] at baseline and 6 months.

2. PRIMARY OUTCOME

2.1 De Paul Symptom Questionnaire – Post-exertional malaise (DSQ-PEM)

The baseline date will be considered as the date of baseline data collection. The DSQ-PEM is a 10-item questionnaire. Questions 1-5 are measured on a five-point Likert scale with a ‘frequency’ domain (0 = none of the time, 1 = a little of the time, 2 = about half the time, 3 = most of the time, and 4 = all of the time) and a ‘severity’ domain. (0 = symptom not present, 1 = mild, 2 = moderate, 3 = severe, and 4= very severe). Questions 6-8 and 10 are dichotomous yes/no responses, and question 9 asked ‘if you feel worse after activities, how long does this last?’ with six options: ≤1 h, 2-3 h, 4-10 h, 11-13 h, 14-23 h, or ≥24 h. The DSQ-PEM sum is the sum of questions 1-5 (frequency and severity), expressed out of 100.

3. SECONDARY OUTCOMES

3.1 Patient health questionnaire (PHQ-4)

The baseline date will be considered as the date of baseline data collection. The PHQ-4 is a 4-item questionnaire (Kroenke, Spitzer, Williams, & Lowe, 2009). Questions 1-4 are measured on a four-point Likert scale with 0= not at all, 1= several days, 2=more than half the days, and 3=nearly every day. The PHQ-4 sum is the scores of each of the 4 items. Scores are rated as normal (0-2), mild (3-5), moderate (6-8) and severe (9-12). Total score ≥3 for first 2 questions suggests anxiety. Total score ≥3 for last 2 questions suggests depression.

3.2 Fatigue severity scale (FSS-7)

The baseline date will be considered as the date of baseline data collection. The FSS-7 is a 7-item questionnaire that measures the impact of fatigue (Krupp et al., 1989). Questions 1-7 are measured on a seven-point Likert scale with 1 (strongly disagree), 4(neither agree nor disagree) and 7 (strongly agree). A visual analogue scale is also included with the scale; respondents are asked to denote the severity of their fatigue over the past 2 weeks by placing a mark on a line extending from “no fatigue” to “fatigue as bad as could be.” Higher scores on the scale are indicative of more severe fatigue.

3.3 12-Item Short Form Survey (SF-12)

The baseline date will be considered as the date of baseline data collection. The SF-12 is a 12-item questionnaire that measures self-reported health-related quality of life covering physical (PCS) and mental health (MCS) domains (Ware, Kosinski, & Keller, 1996). Question 1 is measured on a five-point Likert scale with 1=excellent, 2= very good, 3=good, 4=fair and 5=poor. Questions 2-3 are measured on a three-point Likert scale with 1=yes, limited a lot, 2=yes, limited a little, and 3= no, not limited at all. Questions 4-7 are measured are dichotomous yes/no responses. Question 8 is measured on a five-point Likert scale with 1=not at all, 2=a little bit, 3= moderately, 4=quite a bit, and 5=extremely. Questions 9-11 are measured on a six-point Likert scale with 1= all of the time, 2=most of the time, 3=a good bit of the time, 4=some of the time, 5=a little of the time, and 6=none if the time. Question 12 is measured on a five-point Likert scale with 1=all of the time, 2=most of the time, 3=some of the time,4=a little of the time, and 5=none of the time. Scores above 50 indicate a better-than-average health-related quality of life, while scores below 50 suggest below-average health.

3.4 EuroQol- 5 Dimension (EQ5D)

The baseline date will be considered as the date of baseline data collection. The study uses the EQ-5D-5L version to assess health status and produces a single index value for health status for use in the calculation of quality-adjusted life years to inform health economics evaluation of investigative interventions [10]. The instrument consists of an EQ-5D-5L descriptive system and an EQ-5D-5L visual analogue scale. The descriptive system has 5 dimensions assessing mobility, self-care, usual activity, pain/discomfort, and anxiety. Each of these dimensions has 5 levels of severity which participants are asked to select one of them to best describe their health status ‘today’: no problems, slight problems, moderate problems, severe problems, and extreme problems. Based on participants’ responses from these 5 dimensions, a single index value will be calculated as detailed by Devlin et al [10]. The single index values are on a scale of 0 (full health) to 1 (state equivalent to dead) and health states considered to be worse than dead attain negative values (<0).

3.5 General self-efficacy scale (GSE)

The baseline date will be considered as the date of baseline data collection. The GSE is a 10-item questionnaire to assess self-reported self-efficacy (Schwarzer, & Jerusalem, 1995). Questions 1-10 are measured on a four-point Likert scale with 1=not at all true, 2=hardly true, 3=moderately true, and 4=exactly true. The total score is calculated by finding the sum of all items. Total score ranges between 10 and 40 with a higher score indicating more self-efficacy.

3.6 Breathlessness (MRC Dypsnoea scale)

The baseline date will be considered as the date of baseline data collection. The MRC Dyspnoea scale is a 5-item questionnaire that assess the degree of baseline functional disability due to dyspnoea (Mahler, & Wellis, 1988). Questions 1-5 are measured on a four-grade scale with 0= I get breathless with strenuous exercise, 1=I get short of breath when hurrying on level ground or walking up a slight hill, 2=On level ground, I walk slower than people of my age because of breathlessness, or I have to stop for breath when walking at my own pace on the level, 3=I stop for breath after walking about 100 yards or after a few minutes on level ground, and 4= I am too breathes to leave the house or I am breathless when dressing/undressing. Total score ranges between 0 and 12 with lower scores indicating worse severity of dyspnoea.

3.7 Cognitive function; Smartphone-based symbol digit modalities test (SDMT)

The baseline date will be considered as the date of baseline data collection. The smartphone-based SDMT (Pham et al., 2021), is a smartphone adaptation of the cognitive test, the symbol-digit modalities test (SDMT) examines processing speed and sustained attention by primarily assessing complex visual scanning and tracking. The test compromises of pairing specific numbers with given geometric figures. Responses are given by pressing correct option on the phone display. Total raw score is calculated as number of correct responses to the total number of all responses given in 90 seconds interval.

3.8 Pain Visual Analogue Scale (VAS)

The baseline date will be considered as the date of baseline data collection. The VAS is a validated pain rating scale first developed by Hayes and Patterson (1921), and scores are recorded by dragging a mark on a 10-cm line that represents a continuum between ‘no pain’ and ‘worst pain’. The findings suggested that 100-mm VAS ratings of 0 to 4mm can be considered no pain, 5 to 44 mm, mild pain; 45 to 74 mm, moderate pain; and 75 to 100 mm, severe pain.

3.9 The Edinburgh Neurosymptoms Questionnaire (ENS)

The baseline date will be considered as the date of baseline data collection. ENS is a 30-item yes/no survey which include the addition of 241 yes/no sub-questions designed to assess the presence and nature of: blackouts, weakness, hemisensory syndrome, memory problems, tremor, pain, fatigue, globus, multiple medical problems, and operations (Shipston-Sharman et al., 2018).

3.10 The Symptom Questionnaire (SQ)

The baseline date will be considered as the date of baseline data collection. The SQ-48 is a 92-item yes/true/no/false questionnaire with brief and simple items state scales of depression, anxiety, anger-hostility, and somatic symptom (Kellner, 1987). Symptom subscales are added together and scored 1 when the answer is YES/TRUE.

4. ANALYSIS OBJECTIVES

The primary trial objective is to determine efficacy of adaptive pacing to decrease DSQ-PEM sum score from month 0 to month 6. This will be addressed with a mixed (between and within) effects analysis of variance (ANOVA).

5. SAMPLE SIZE

To determine sample size, our primary outcome variable was the DSQ-PEM. Using previous work, a minimum clinically relevant difference can be estimated as a change of 13 points on a 100-point scale [9]. Assuming a standard deviation (SD) of 25 [9], this resulted in a pairwise effect size of d=0.5 (Cohen’s f=0.25). We calculated our desired sample size for a two-way mixed-model (within- and between-subjects) analysis of variance (ANOVA). Using the WebPower package in R Studio, and the wp.rmanova function, with two groups, two time points, a medium effect size (f=0.25), assuming sphericity, an alpha of 0.05, desired statistical power of 0.9, testing for an interaction effect, the total n was 170 (85 per group). Consequently, we aimed to recruit 125 participants per group to allow for 30% drop-out.

6. RANDOMISATION, BLINDING, AND OUTCOME ASSESSMENT

Participants will be randomised to one of the two trial arms using 1:1 allocation ratio.

Randomisation will be performed by a web-based online randomisation system (Study Randomizer). We will randomise participants remotely and blinding is impossible given the nature of the intervention. This is an unblinded study recognising that participants cannot be blinded to their treatment allocation. The outcome assessor (LH) will be blinded to the group coding ("1" or "2") and will have no interaction with participants prior to conclusion of the trial. Outcome assessors will not contact participants as data are collected remotely via a mobile app. The PI (NS) will be responsible for 'pulling down' the data from Firebase and assembling the outcome data in a .csv file, with groups coded as "1" or "2". No conversations between the outcome assessor and participants will take place. The intention is that the same assessor will carry out all outcome assessments for consistency, but if unblinding has occurred then an alternative assessor will be used as blinded assessments will take priority over assessments by the same assessor.

7. DATA SOURCES

The data used in this study will come from self-reported questionnaires collected using a mobile app. Data will be stored on the Firebase database platform with the exception of the randomisation list which is held on the post-doc's secure personal computer. Electronic data will be extracted from the system at regular intervals in order to facilitate validation of the data and monitoring of the trial progress. Any spurious data will be queried and checked for consistency with data management before data lock. Personal records will not be accessible to research staff.

8. PROTOCOL NON-COMPLIANCES

For the purposes of the analyses, participants who are deemed not to have adhered to the data entry requirements (i.e. completed the DSQ-PEM at month 0 and month 6) will be considered as non-compliances. Participants allocated to usual care may purchase their own wearable or engage in adaptive pacing of their own accord. However, developing the framework to deliver just-in-time messages and data repackaging is complex, tailored to individual requirements, and not possible with 'off the shelf' wearables. Therefore, even if participants were to acquire a wearable device or engage in adaptive pacing, they would not have access to bespoke just-in-time-adaptive intervention support.

9. ANALYSIS POPULATIONS

The per protocol analysis set will be used as the primary set for analysis. This includes:
1) all participants for whom consent is obtained and;
2) treatment assignment as per randomised list regardless of 'circumstances' after randomisation, and;
3) all participants with primary outcome data at month 0 and month 6, and;
4) only experimental group participants who wore the wearable throughout the intervention period (with the exception of sleeping and charging).

Note that this analysis set for baseline tables of demographics and characteristics of participants illustrated in tables will be primarily based on participants with DSQ-PEM data (primary outcome data) at 6 months and at baseline. Regarding baseline missing data for this study, the first assessment of the primary outcome is at 0 months, guided by the post-doc, so minimal data should be missing at this point. The objective is to explore the effectiveness of the intervention among participants who adhered to key components of this intervention as intended.

10. STATISTICAL METHODOLOGY

The primary outcome (DSQ-PEM), and secondary outcomes will be pre-treatment (baseline), and at 6 months post-baseline. Any adverse events may be reported at any time during the study period.

10.1 Demographics and baseline characteristics

Demographics and baseline characteristics of participants will be summarised within each treatment arm but not assessed for baseline comparability Because of randomisation, we did not undertake analysis of baseline equivalence, since the null hypothesis must be true and any differences due to chance [11]. The per participant denominator is the number of participants with month 0- and month 6-month primary outcome data. Because of the nature of this remote intervention, only age and gender will be collected as baseline characteristics, given the logistical challenge, participant burden, and cost implication of participant travel. For continuous variables, summaries will comprise the number of participants and either of the following depending on the distribution of the data: 1) Mean (95% confidence intervals) or; 2) Median (interquartile range). Summaries for categorical outcomes will comprise the number of participants and their respective proportion as a percentage in that category. The results will be presented as shown in Table 1.

10.2 Recruitment and data completeness

The following summaries will be presented for all participants screened for entry to the study, by identification or recruitment source and overall. For the purpose of recruitment, the following summaries will be collected: 1) The number of participants screened, 2) The number of participants recruited, 3) Number and percentage of participants not recruited and the reasons for non-recruitment. Relevant summaries on recruitment, consent and data completeness during follow-up will be presented in a CONSORT flowchart [11]. Reasons for withdrawal at different follow-up times will also be summarised by treatment arm.

10.3 Dealing with deaths

In this trial population, few deaths during the trial are expected. We discussed how to handle deaths during analysis. There was agreement that the influence of the intervention on mortality such as increasing the risk of mortality is very unlikely. In addition, the interpretation of imputed data for this study, such as PEM for participants who have died is difficult. In this regard, missing data due to deaths will not be imputed. Thus, deaths prior will be excluded in any analysis. The number of individuals who have died during the trials will be reported by treatment group and presented in the CONSORT flowchart.

10.4 Differential characteristics of completers versus non-completers at baseline

This section aims to explore whether completers differ systematically at baseline from non-completers with respect to their key characteristics. Completers are participants with primary outcome data at 0 and 6 months and non-completers are those with missing primary outcome data for any reason, excluding death. Exploring the patterns of missing data is important to aid interpretation of results. First, the reasons for missing data will be summarised appropriately. Second, descriptive statistics will be used to further this objective. Participants who have died prior to 6 months follow-up assessment will be excluded in this exploratory analysis since they will not be included in any subsequent analysis. Third, baseline variables associated with the primary endpoints will be descriptively explored among those with available data. For example, using scatterplots of primary endpoints against baseline variable stratified by the treatment allocation.

11. ANALYSIS OF PRIMARY AND SECONDARY ENDPOINTS AT 6 MONTHS

The primary endpoints assessed following 6 months of treatment (from baseline) will be the change in PEM frequency and severity, measured using the sum of questions 1-5 in the DSQ-PEM. The primary analysis will be based on per protocol analysis. The denominators may differ across endpoints depending on data completeness at 6 months. For the change in DSQ-PEM sum at 6 months from baseline, the measure of intervention effect will be the mean difference in change in DSQ-PEM sum between the adaptive pacing and standard care groups. The primary analyses will utilise a mixed-effects (within and between groups) analysis of variance (ANOVA). Results will be reported and presented as means (95% confidence intervals) at baseline and post-intervention in both groups. Associated P-values, and measures of effect size for within-group (i.e. time) effects, between-group (i.e. treatment allocation) effects, and interaction effects (time x allocation) will be reported.

In the case of categorical variables (DSQ-PEM questions 6-10) we will use McNemar's Test for paired samples (pre- to post- intervention), or Chi squared test for between group effects (intervention vs. control).

12. EXPLORATORY ANALYSIS OF RESPONSES BY MONTH

This section aims to provide granularity in terms of temporal symptom change. If a pre- and post-intervention analysis suggests similarity between groups, it is still possible one group reduced their symptoms at a faster rate. Therefore, to examine the effects of time and group on the primary and secondary endpoints, we will use the following linear mixed-effects model:

$$Outcome_{in} = \beta_0+\beta_1\cdot \mathrm{time}_{in} +\beta_2\cdot \mathrm{group}_{in}+\beta_3\cdot(\mathrm{time}_{in}\cdot \mathrm{group}_{in})+(1|\mathrm{Subject})+ \varepsilon_i$$

All domains of endpoints represented the repeated-measures *outcome* for subject$_{in}$ and serve as outcome measures whereas *time* (continuous variable with 7 levels [consecutive months]), *group* (categorical variable with 2 levels [intervention and control]) will be modelled as predictors and treated as fixed effects alongside their two-way pairwise interactions. Moreover, random effects will be assumed for *participants*, with random slopes per the predictor time introduced in the model if this addition does not result in a convergence error. If we assume data will be missing at random, and linear mixed effects models handle missing data without requiring imputation [12]. Estimated marginal means and 95% confidence intervals will be calculated alongside comparisons made using post-hoc Holm-Bonferroni adjustments.

13. REFERENCES

1. Hayes LD, Ingram J, Sculthorpe NF. More than 100 persistent symptoms of SARS-CoV-2 (long COVID): a scoping review. Frontiers in Medicine. 2021;8.

2. Bayliss K, Goodall M, Chisholm A, Fordham B, Chew-Graham C, Riste L, et al. Overcoming the barriers to the diagnosis and management of chronic fatigue syndrome/ME in primary care: a meta synthesis of qualitative studies. BMC Fam Pract. 2014;15:44.

3. Deumer U-S, Varesi A, Floris V, Savioli G, Mantovani E, López-Carrasco P, et al. Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS): An Overview. Journal of Clinical Medicine. 2021;10:4786.

4. Sanal-Hayes NEM, Mclaughlin M, Hayes LD, Mair JL, Ormerod J, Carless D, et al. A scoping review of 'Pacing' for management of Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS): lessons learned for the long COVID pandemic. Journal of Translational Medicine. 2023;21:720.

5. Goudsmit EM, Howes S. Pacing: A strategy to improve energy management in chronic

fatigue syndrome. Health Psychology Update. 2008;17:46.

6. Hardeman W, Houghton J, Lane K, Jones A, Naughton F. A systematic review of just-in-time adaptive interventions (JITAIs) to promote physical activity. Int J Behav Nutr Phys Act. 2019;16:31.

7. Perski O, Hébert ET, Naughton F, Hekler EB, Brown J, Businelle MS. Technology-mediated just-in-time adaptive interventions (JITAIs) to reduce harmful substance use: a systematic review. Addiction. 2022;117:1220–41.

8. Cotler J, Holtzman C, Dudun C, Jason LA. A Brief Questionnaire to Assess Post-Exertional Malaise. Diagnostics. 2018;8:66.

9. Jason L, Ohanian D, Brown A, Sunnquist M, McManimen S, Klebek L, et al. Differentiating Multiple Sclerosis from Myalgic Encephalomyelitis and Chronic Fatigue Syndrome. Insights Biomed. 2017;2:11.

10. Devlin NJ, Shah KK, Feng Y, Mulhern B, van Hout B. Valuing health-related quality of life: An EQ-5D-5L value set for England. Health Econ. 2018;27:7–22.

11. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, et al. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c869.

12. Gabrio A, Plumpton C, Banerjee S, Leurent B. Linear mixed models to handle missing at random data in trial-based economic evaluations. Health Economics. 2022;31:1276–87.