



Statistical Analysis Plan

M-FIT

Version 0.1

April 11, 2022

Sponsor:	The University of Queensland acting through the Australasian Kidney Trials Network (AKTN)
Registration (ANZCTR):	ACTRN12620000408987
Chief Investigator:	Professor Allison Tong

Contents

Version History	4
Preface	5
Abbreviations	6
1 Introduction	7
1.1 Background and rationale	7
1.2 Study Objectives	9
1.2.1 Primary Objective	9
1.2.2 Secondary Objectives	9
2 Study Design	10
2.1 Overview	10
2.2 Target Population	10
2.2.1 Inclusion Criteria	10
2.2.2 Exclusion Criteria	11
2.3 Interventions	11
2.4 Randomisation	12
2.5 Blinding	13
2.6 Sample Size	13
3 Study Outcomes	14
3.1 Primary Outcome	14
3.2 Secondary Outcomes	14
3.2.1 Patient-reported outcomes	14
3.2.2 Exercise-physiologist assessed outcomes	17
3.2.3 SONG-HD outcomes	18
3.2.4 Safety outcomes	19
3.2.5 Adherence	20
3.2.6 Economic outcomes	20
3.3 Baseline demographics	20
3.4 Subgroups	21
4 Analysis - General	22
4.1 Participant Flow	22
4.2 Descriptive Statistics	22
4.3 Intervention Adherence	22
4.4 Analysis Sets	23
4.4.1 Intention-to-treat Analyses	23
4.4.2 Per-Protocol Analyses	23

4.5	Comparisons	24
5	Statistical Modelling	25
5.1	Primary Outcome Model	25
5.1.1	Model specification	25
5.2	Subgroup Analyses	27
5.3	Secondary Analyses of Primary Outcome	28
5.4	Secondary Outcome Analyses	28
5.4.1	Patient-reported Outcomes	28
5.4.2	Exercise professional assessed outcomes	31
5.4.3	Clinical Outcomes	32
5.5	Model Checks	33
5.6	Qualitative Analyses	34
5.7	Missing Data	34
5.7.1	Baseline	34
5.7.2	Follow-up	35
5.8	Sensitivity Analyses	35
5.9	Safety Analyses	36
5.10	Exploratory Analyses	36
6	Statistical Quantities	37
6.1	Comparison with best and control	37
6.2	Intervention superiority	37
6.3	Intervention inferiority	38
6.4	Intervention effectiveness	38
6.5	Intervention ineffectiveness	38
7	Trial Adaptations and Statistical Decisions	39
7.1	Sequential Analyses	39
7.2	Stopping rules	39
7.3	Response-Adaptive Randomisation	40
8	Interim Analyses and Trial Reporting	42
8.1	Trial Adaptations	42
8.2	Trial Reporting	42
	References	43
	Appendix I: Trial Operating Characteristics	45
	Appendix II: Study Database and Derived Variables	49

Version History

Version	Date	Author	Description
0.1	2021-04-23	MAJ	First draft
0.2	2021-07-27	JAT	Revisions
0.3	2021-09-23	JAT	Revisions of outcomes
0.4	2021-10-21	JAT	Revisions of section 6 and 7
0.5	2022-03-17	JAT	Revise primary model

Preface

This statistical analysis plan (SAP) outline the data and procedures for analysing effectiveness of trial arms from the protocol M-FIT: an adaptive, randomised trial examining the comparative effectiveness of structured exercise programs on fatigue experienced by adult patients receiving dialysis.

The following documents were reviewed when preparing this SAP:

- M-FIT Study Protocol version 0.9.7, 2 September 2021
- M-FIT Data Dictionary, 19 October 2021 (REDCap extract date)

Abbreviations

Abbreviation	Definition
30STS	30-second sit-to-stand test
AKTN	Australasian Kidney Trials Network
CKD	Chronic Kidney Disease
FACIT-Fatigue	Functional Assessment of Chronic Illness Therapy Fatigue
HADS	Hospital Anxiety and Depression Scale
HRQoL	Health-Related Quality of Life
HD	Haemodialysis
PD	Peritoneal dialysis
M-FIT	Structured exercise program to reduce fatigue in patients receiving dialysis: an adaptive trial
PFS	Pittsburgh Fatigability Scale
POMA	Performance Oriented Mobility Assessment
PROMIS	Patient-Reported Outcomes Measurement Information System
PROMIS-APSRA	Patient-Reported Outcomes Measurement Information System - Ability to Participate in Social Roles and Activities
QALY	Quality of life years
SAP	Statistical analysis plan
SONG	Standardised Outcomes In Nephrology
SONG-HD	Standardised Outcomes In Nephrology - Haemodialysis
SONG-HD-F	Standardised Outcomes In Nephrology - Haemodialysis - Fatigue

1 Introduction

1.1 Background and rationale

Fatigue is one of the most common and debilitating symptoms in patients on dialysis. Patients with chronic kidney disease (CKD) have also explicitly identified non-pharmacological self-management of lifestyle interventions to improve symptoms as a top research priority.

Based on some evidence to indicate that regular exercise improves cardiovascular outcomes, physical activity and quality of life across all stages of CKD, the primary objective of the M-FIT (Structured exercise program to reduce Fatigue In patients receiving dialysis: an adaptive Trial) study is to determine whether home-based exercise can improve fatigue in patients on dialysis.

M-FIT is a multi-centre, adaptive randomised comparative effectiveness trial of four structured exercise programs in adult patients receiving dialysis. The trial will compare exercise regimes delivered by exercise physiologists and a mobile-app that aim to alleviate fatigue in patients on dialysis. Specifically, the trial will evaluate

- the comparative effectiveness of a structured exercise program on fatigue in patients receiving dialysis
- the effectiveness of each therapy relative to an attention-control arm.

The study is open-label; patients, clinicians and analysts will know who was allocated to which therapy. However, trial interim results are confidential until the final analysis.

The primary analysis is based on the 13-item 52-point FACIT-Fatigue¹ (Functional Assessment of Chronic Illness Therapy Fatigue) score that will track individual-level fatigue over time. The score ranges from 0 to 52 and is reverse coded so that higher scores imply less fatigue. The intervention therapies used as study arms comprise:

1. attention-control: non-exercise based activity (stretching)
2. walking
3. resistance-only
4. resistance and aerobic

¹FACIT-F (as opposed to FACIT-Fatigue) is a related 40-item measure that also assesses self-reported fatigue and its impact upon daily activities and function. M-FIT uses the 13-item scale.

The therapies are delivered via an initial face-to-face assessment with an exercise physiologist and are followed-up over 12 weeks with weekly telehealth visits in weeks 1 to 4, 6 and 8 and a final follow up session in week 12. Additionally, there is direction from a mobile-app.

Data collection is in part entered by study personnel, but also provided directly by the participant since they respond to surveys via the mobile application.

1.2 Study Objectives

1.2.1 Primary Objective

To assess the comparative effectiveness of alternative structured exercise programs, co-designed with patients, on fatigue, as measured by FACIT-Fatigue, in adult patients receiving dialysis.

1.2.2 Secondary Objectives

1. To assess the comparative effectiveness of four structured exercise programs on additional measures of fatigue, quality of life, mood, life participation, frailty, changes in body composition, exercise capacity, neuromuscular fitness, balance, physical activity, sleep, exercise adherence, SONG core outcomes, and hospital admissions.
2. To compare the cost-effectiveness of four structured exercise programs in terms of incremental cost, and incremental health outcomes (quality-adjusted life year (QALY) and clinically important difference in fatigue).
3. To assess the impact, fidelity, facilitators, and barriers of implementing the exercise programs in patients receiving dialysis.
4. To evaluate the influence (if any) of participant preferences for the assigned interventions on fatigue and exercise adherence.
5. To assess hospital admissions, mortality, exercise-related injuries, and cardiovascular events for all participants.

2 Study Design

The following information is summarised here from the study protocol. For full details, refer to the relevant section in the study protocol.

2.1 Overview

The study is a prospective, multi-centre, adaptive, randomised comparative effectiveness trial. The design uses group sequential methods and Bayesian response-adaptive randomisation.

2.2 Target Population

All patients on haemodialysis (HD) or peritoneal dialysis (PD) who can provide informed consent in English, will be invited to participate in the trial. Patients who are new to dialysis or those who have received a kidney transplant will be excluded. Patient-reported outcome measures form a substantial part of the assessment in the trial, and only English versions are used. Therefore, the eligibility criteria ensure that participants can comprehend the questions and respond accurately. To ensure safety of the participants in the trial, exercise professionals will screen eligible patients for their capability to carry out movements required in all four exercise arms.

2.2.1 Inclusion Criteria

To be eligible to participate in this trial, the participant must satisfy the following criteria:

1. on maintenance haemodialysis or peritoneal dialysis (>3 months) with a life expectancy of ≥ 12 months.
2. > 18 years of age
3. able to provide informed consent
4. able to speak, read and write English
5. access to a smart phone or tablet
6. physically capable of carrying out all three exercises at intensity level 1 at minimum (as assessed by the site exercise professional).

2.2.2 Exclusion Criteria

Participants are excluded from M-FIT if they meet any of the following criteria:

1. presence of known cardiovascular disease that places the participant at an unacceptable risk of untoward event occurring during exercise training (as deemed by treating physician)
2. have received or are expected to receive a kidney transplant within 12 months
3. currently meeting the physical activity guidelines (150 mins/week of moderate intensity aerobic (cardio) activity and 2 sessions/week of resistance training)

2.3 Interventions

The trial interventions consist of three different structured exercise programs: walking, resistance training, combination aerobic (cardio) and resistance training, and a single attention-placebo control group (stretching). There are five levels of exercise prescription within the walking, resistance training, and aerobic and resistance training intervention arms. At baseline, the administering exercise professional will select the level of exercise deemed most appropriate for the patient. The control arm will be assigned a set of stretches to complete for 12 weeks.

All participants will be provided with a mobile application. The mobile application gives the participant access to demonstration videos applicable to their allocated group, and trial outcome assessment questionnaires. Participants will be asked to download the application to their device. The site exercise professional and study coordinator will be able to help with set up and setting of the initial treatment arm and exercise intensity level. Once a participant has finished the treatment portion of the trial (12 weeks) they will be able to access all treatment arms and intensity levels.

The complete intervention will consist of:

1. At baseline: an assessment (90 mins) with an exercise professional to assign a suitable initial intensity level and demonstration of the exercises, provision of the app and instructions by the research team.
2. Weeks 1 to 4: weekly check-up sessions (10 min telehealth or in person) with the exercise professional and possible adjustment of the intensity level based on clinical and personal factors to ensure appropriate exercise intensity.

3. Weeks 5 to 8: fortnightly check-up sessions (10 min telehealth or in person) with the exercise professional and possible adjustment of the intensity level.
4. Weeks 12: final assessment session (90 mins) which repeats baseline assessments to assess exercise capacity. At this final assessment the exercise professional will include motivational interview and adherence strategies to improve exercise adherence during the 12 to 36-week period.
5. Week 36: follow-up session (90 mins) to assess exercise capacity.

Participants' adherence to the interventions will be assessed with a self-report measure within the M-FIT application. Research coordinators for each site will monitor and document participants' data entry on adherence on a weekly basis with a check-up call to participants who have missed two or more sessions.

For full details of what each individual intervention entails, refer to the study protocol.

2.4 Randomisation

Eligible participants will be randomised by research staff via a proprietary randomisation system and their assignment transcribed to the RedCap patient database. The randomisation system was developed, validated and will be hosted by University of Sydney and has secure access, full audit trail and redundancy. Participants, clinical staff and analysts are unblinded to allocation.

Participants will initially be randomised in 1:1:1:1 allocation between the four intervention arms. Following each interim analysis, the allocation ratios to intervention arms will be updated. The target allocation to the attention-control arm will remain fixed as the reciprocal of the number of arm still available in the trial, for example, if one arm is dropped following an interim, then the allocation to control would be 1/3. For the exercise arms, the target allocation will be proportional to the posterior probability that the intervention is most effective at reducing fatigue at 12 weeks.

If an intervention's posterior probability of being best falls below a pre-specified threshold then the allocation to that intervention may be set to 0 and the intervention dropped from the trial. If the probability exceeds a pre-specified threshold then that intervention may be declared superior and randomisation stopped.

Randomisation probabilities will be updated by the trial statistician directly after the completion of each interim analysis (**or after meeting with the DSMB?**). [Section 7.3](#)

contains further details on the calculation used to compute the randomisation probabilities.

2.5 Blinding

Due to the nature of the intervention, the site investigators, AKTN, treating clinicians, patients and the analysis group will not be blinded to treatment allocation. The trial steering committee, statistical analysis plan developers, and outcome assessors will be blinded.

2.6 Sample Size

The study sample size was selected on the basis of feasibility and to target a mean difference of half a standard-deviation in FACIT-Fatigue scores. Assuming a two-sample independent one-sided t -test of size 0.025 and an alternative hypothesis of a difference in means of $\Delta = 0.5\sigma$ (σ the standard deviation), a sample size of 85 per arm has power 0.9. Ignoring multiplicities, this results in a total sample size of 340 across four arms. Allowing for a conservative amount of drop-out, this total sample size was increased to 400.

Further evaluation of power and trial operating characteristics for the adaptive design were accomplished via simulations. We undertook 2,000 simulations under each of the trial scenarios of no treatment effects, and alternative scenarios where any superior treatments had an effect size of 0.5σ on FACIT-Fatigue scores, with a maximum sample size of 400 and assumed 20% probability of drop-out. The target allocation to each arm was initially equal and updated by using Bayesian response adaptive randomisation. The simulations assumed compound symmetric correlation of 0.25 between the longitudinal outcomes of a participant.

In the null scenario, using a superiority decision threshold of 0.98 and inferiority threshold of $0.02/(\text{number of active arms} - 1)$, the probability of deciding any arm superior was $< 1\%$ and the probability of dropping each arm as inferior was approximately 10%. When one intervention was superior to all others by 0.5σ , a superiority threshold of 0.98 resulted in a decision of superiority of the superior arm in 87% of trials and a decision of inferiority for the superior arm in 0% of trials.

3 Study Outcomes

A summary of the study outcomes and their measurement occasions are provided in Table 1. For full details of each outcome refer to the relevant section. Visits at baseline, week 12 and week 36 occur in clinic, and contact at week 4 and 8 are via telehealth.

All patient-reported outcomes, including adherence questionnaires, will be collected via REDCap surveys embedded into the M-FIT mobile application.

3.1 Primary Outcome

The primary outcome is fatigue as measured by the **Functional Assessment of Chronic Illness Therapy - Fatigue Scale (FACIT-Fatigue)** [1, 2] at 12-weeks post-randomisation. FACIT-Fatigue is a 13-item Likert scale, with each Likert item scored from 0 to 4. Level of fatigue is measured as the total score aggregated across the 13 items, ranging from 0 to 52. Lower scores imply a higher level of fatigue.

If items are missing, the score is calculated using the average of the items with responses. If fewer than 50% (i.e. 7) items have a response, then the score is missing.

FACIT-Fatigue scores will be measured at baseline visit (week 0) and study visits at weeks 4, 8, 12, and 36. The data must be collected within five days of the scheduled visit. For FACIT-Fatigue example norms for males and females between ages 18 to 70+ see [3]. For an example in a dialysis population see [4].

3.2 Secondary Outcomes

3.2.1 Patient-reported outcomes

The **Standardised Outcomes In Nephrology - Haemodialysis - Fatigue (SONG-HD-F)** [5] is a 3 item Likert scale designed to measure fatigue in the past week. Each Likert item is a scale from 0 to 3 resulting in an aggregated score ranging from 0 to 9.

This outcome is reported at weeks 0, 4, 8, 12, and 36.

The **Pittsburgh Fatigability Scale (PFS)** [6] is a one-page, self-administered questionnaire that asks about level of exertion on a scale from 0 to 5 on 10 different activities.

3 Study Outcomes

Outcome	Baseline (Clinic)	Week 4 (Phone)	Week 8 (Phone)	Week 12 (Clinic)	Week 36 (Clinic)
Primary outcome					
FACIT-Fatigue	✓	✓	✓	✓	✓
Secondary outcomes					
<i>Patient-reported outcomes</i>					
SONG-HD-F	✓	✓	✓	✓	✓
PFS	✓	✓	✓	✓	✓
EQ-5D-5L	✓	✓	✓	✓	✓
HADS	✓	✓	✓	✓	✓
PROMIS-APSRA	✓	✓	✓	✓	✓
mSUS					✓
<i>Exercise-professional assessments</i>					
30 sec sit to stand	✓			✓	✓
Modified wall push-up	✓			✓	✓
Timed up-and-go	✓			✓	✓
6 min walk test	✓			✓	✓
Arm curl test	✓			✓	✓
Hand grip strength	✓			✓	✓
Tinetti POMA	✓			✓	✓
Fried Frailty Index	✓			✓	✓
BMI	✓			✓	✓
Waist circumference	✓			✓	✓
Physical activity/sleep (actigraph)	✓			✓	✓
<i>SONG core outcomes</i>					
Vascular access function					✓
Technique survival					✓
Peritoneal dialysis infections					✓
<i>Adverse events of interest</i>					
Mortality		✓	✓	✓	✓
Hospitalisations		✓	✓	✓	✓
Cardiac events		✓	✓	✓	✓
Falls		✓	✓	✓	✓
Exercise related injuries		✓	✓	✓	✓
<i>Exercise adherence¹</i>					
Pre-exercise safety questionnaire		✓	✓	✓	✓
Post-exercise questionnaire		✓	✓	✓	

¹ After every exercise session.

Table 1: Summary of study outcomes and their measurement occasions.

For each activity, a response for both physical and mental fatigue are elicited. A total score for physical and mental fatigue are obtained by summing the respective rating for each activity, resulting in two scores each ranging from 0 to 50.

This outcome is reported at weeks 0, 4, 8, 12, and 36.

The **EuroQol 5 Dimension 5 Level (EQ-5D-5L)** is a standardised self-report survey measure of health-related quality of life (HRQoL) that comprises five dimensions: mobility, self-care, usual activities, pain and discomfort, and anxiety and depression. The five dimension each have five response levels of severity giving health scores ranging from 11111 (full health) to 55555 (worst health), equating to $5^5 = 3125$ potential health states. For analysis, the responses are routinely converted into an index representing the utility of the health state. The conversion, while not recommended by the EuroQol Group, proceeds by applying a societal preference function which generates scores anchored at 0 for death and 1 for perfect health and represent a societal valuation of health-related quality-of-life. The conversion only works for complete data. The tool also includes a visual analog scale (VAS) which records the respondent's self-rated health status graded on a 0-100 scale, where higher scores imply higher HRQoL.

This outcome is reported at weeks 0, 4, 8, 12, and 36.

See [7] and [8] for more details on methods of analysis.

The **Hospital Anxiety and Depression Scale (HADS)** [9] is a questionnaire designed to measure anxiety and depression in a general medical population of patients. It comprises seven questions each for anxiety and depression scored separately. Each item is scored from 0 to 3 and aggregated across the seven questions resulting in two scores in the range of 0 to 21, one for anxiety and one for depression. This outcome is reported at weeks 0, 4, 8, 12, and 36.

The **Patient Reported Outcomes Measurement Information System - Ability to Participate in Social Roles and Activities (PROMIS-APSRA) v2.0 4a** [10] is a 4 item questionnaire designed to assess the perceived ability to perform one's usual social roles and activities. The 4 Likert items are on a scale of 1 to 5 resulting in aggregated raw scores ranging from 4 to 20. This outcome is reported at weeks 0, 4, 8, 12, and 36.

The **modified System Usability Scale (mSUS)** assesses the experience (user friendliness, complexity, consistency, and user confidence) that users have with systems (applications, hardware) as well as benchmarking products against well-known, highly used systems. The survey consists of three sections. The first uses 10 items graded on a 5-point scale to assess usability. The second section requests an overall rating of user-friendliness on a 7-point scale. The third requests a letter grade rating of user-friendliness.

This item is self-reported at week 36 follow-up session.

3.2.2 Exercise-physiologist assessed outcomes

The **30 sec sit to stand test (30STS)** [11, 12, 13] is a test to assess lower body strength, endurance, and coordination. The patient completes as many stands from a chair as possible in 30 seconds.

The **modified wall push-up test (MWP)** is aimed at testing a patients upper body strength and endurance. The patient completes as many push-ups as possible in 30 seconds while standing against a wall.

The **timed up-and-go test (TUG)** [14, 15] is used to assess mobility, balance, walking ability, and fall risk. Patients stand from a seated position, walk 3 meters, turn around, walk back, and sit back down. The aim is to complete this in as little time as possible.

The **arm curl test (ACT)** [11, 16] is used as a predictor of elbow flexor strength. Patients complete as many arm curl repetitions as possible in 30 seconds using a designated dumbbell weight.

The **hand-grip strength test (HGS)** [16] is used as a measure of overall body strength. Using a hand grip dynamometer, patients use one hand to apply as much force as possible in a squeezing action. The amount of force is reported in kilograms.

The **six-minute walk test (6MWT)** [17, 18] is a sub-maximal exercise test performed to assess aerobic endurance. For 6 minutes patients walk as far as possible utilising a 15 or 30-metre stretch of uninterrupted pathway as a course. The total distance walked is reported (metres).

The **Tinetti Performance Oriented Mobility Assessment (POMA)** [19, 20] is a battery of assessments aimed at measuring a patients gait and balance abilities. Scoring is done separately for balance (score range 0 to 12) and gait (score range 0 to 16), and the two results combined for a score out of 28. A lower score implies higher risk of falls.

The **Fried Frailty Index (FFI)** [21, 22] is an assessment of physical functioning through both questions and physical assessments. Five criteria comprise the assessment: shrinkage (weight loss), exhaustion, slowness, physical activity, and weakness. Weight loss is either self-reported (or calculated if data is available) over the last year, exhaustion is

determined by two questions from the Geriatric Depression Scale (GDS), physical activity is self-reported weekly activity, slowness is determined by timed up-and-go, and weakness is determined from hand-grip strength. The sum of scores in these categories classifies individuals into three frailty conditions: not frail, pre-frail and frail.

Body composition will be assessed by **waist circumference** (cm) and **Body Mass Index (BMI)** (kg/m^2).

Minutes of **physical activity** and **sleep** will be measured using an Actigraph wrist accelerometer in the 7 days prior to intervention and the 7 days before intervention ends.

3.2.3 SONG-HD outcomes

SONG-HD core outcomes will be collected at week 36.

Vascular access function

The need for any intervention required to enable and maintain the use of a vascular access for haemodialysis. Vascular access interventions include:

- open surgical or endovascular creation/placement of AVF/AVG
- open surgical revision or endovascular intervention of AVG/AVF
- thrombolysis or thrombectomy of AVG/AVF
- ligation or resection of Arteriovenous access
- repair of aneurysm/pseudoaneurysm
- competing/collateral vein ligation
- Fistulogram (Angiogram) +/- Angioplasty +/- Stenting (including inflow artery, body of AVF/AVG, venous outflow, central vein)
- competing/collateral vein embolisation
- superficialisation/transposition
- distal Revascularisation, Interval Ligation (DRIL)
- proximalisation of the Arterial Inflow (PAI)
- revision Using Distal Inflow (RUDI)
- banding
- CVC insertion
- CVC exchange
- fibrin sheath removal/disruption

- CVC removal

Technique survival

TODO

Peritoneal dialysis infections

TODO

3.2.4 Safety outcomes

Adverse Events of Special Interest (AESI) will be recorded as part of regular data collection activities. The following events are of special interest:

- Death
- Hospitalisation
- Cardiovascular events
- Injurious falls and resultant events
- Hypoglycaemia
- Musculoskeletal injuries
- Diabetes related complications

Serious adverse events (SAE) are any events that:

- Led to death
- Led to serious deterioration in the health of the subject, that either resulted in:
 - A life-threatening illness or injury, or
 - A permanent impairment of a body structure or a body function, or
 - In-patient or prolonged hospitalization, or
 - Medical or surgical intervention to prevent life-threatening illness or injury or permanent impairment to a body structure or a body function.

3.2.5 Adherence

A **pre-exercise safety questionnaire** will be administered immediately before the patient begins every exercise session is used to determine the safety of the patient to complete an exercise session. The site exercise professional is also able to track any injuries, medication changes or health issues a patient may be having and adjust exercise prescription as necessary. The questionnaire will ask about the participants capacity to exercise, and they self-report as limited capacity, then additional questions of their condition will be asked (cold/flu symptoms, nausea, muscle/joint/bone problems, etc.).

The **post-exercise questionnaire** is a simple report on patients self-adherence to the exercise regimen and intensity of exercise. This allows the site exercise professional to monitor patient performance and adjust exercise prescription where necessary to remain in the designated target intensity level. The questionnaire will ask whether the participant completed the days exercise, and how difficult they perceived it to be.

3.2.6 Economic outcomes

Health-care utilisation and associated costs will be investigated by linkage to administrative data. Costings will rely on diagnostic related group (DRG) codes or urgency related groups (URG) or equivalent.

3.3 Baseline demographics

The following information will be collected at baseline interview:

- age
- sex
- ethnicity
- country of birth
- primary kidney disease
- co-morbidities including smoking
- dialysis modality
- occupational status
- highest level of education achieved
- marital status

3.4 Subgroups

Pre-specified subgroups are:

- exercise preference

Refer to [5.2](#) for details of planned subgroup analysis.

4 Analysis - General

4.1 Participant Flow

The flow of participants through the trial will be reported following CONSORT. This will include the number of participants:

- randomly assigned
- receiving each study arm
- followed up
- withdrawn
- included in analysis of the primary outcome.

Information on participant screening and ineligible participants will also be reported, as well as summaries of protocol deviations.

4.2 Descriptive Statistics

Descriptive statistics will be calculated and presented by arm and in aggregate for all relevant baseline variables and study outcomes. Categorical variables will be summarised by counts and proportions, including number missing. Continuous variables will be summarised by minimum, maximum, means and standard deviations or (medians and interquartile ranges), and number missing will be reported.

4.3 Intervention Adherence

Participants are asked to complete a pre and post exercise survey. For each intervention, self-reported adherence will be summarised, including the number and proportion of exercise sessions not completed and the reasons for incompleteness.

4.4 Analysis Sets

4.4.1 Intention-to-treat Analyses

The primary analysis will include all participants randomised to an intervention arm regardless of whether intercurrent events disrupted the allocated exercise regime and irrespective of adherence. That is, we will adopt an intention-to-treat (ITT) estimand (de facto treatment-policy) for the analyses with all randomised participants contributing to the analysis such that:

- patients are analysed in the group they were allocated to
- patients that do not receive (either partially or fully) the intervention will be retained
- protocol deviations will not result in automatic exclusion
- patients who die before end of the study will be included in the analysis population until their time of death

4.4.2 Per-Protocol Analyses

Secondary analyses will investigate per-protocol effects where a participants adherence is taken into consideration. These analyses will use the information collected in the post-exercise questionnaire on self-reported adherence on whether exercise was completed on that day.

Assuming that this will be of interest: What is considered a protocol deviation to cause exclusion from a per-protocol set? How will we account for adherence, what aspect of adherence is most relevant? Restrict the per-protocol analysis population to only those deemed to be "adherers"? An "adherer" anyone who completes at least x% (80%) of exercise sessions as expected? Continuous measure of adherence, e.g. proportion of sessions completed out of total expected number of exercise sessions? What happens post-non adherence? Use adherers data for imputation of missing outcomes?

4.5 Comparisons

For all outcomes, the primary analysis will use the intention-to-treat population with particular interest in the comparison at 12 weeks post-randomisation. Each exercise intervention will be compared to the attention-control arm and against all other exercise interventions. If an exercise arm is dropped from the study the primary comparison will be with respect to participants randomised prior to the decision to drop the arm (concurrently randomised participants). However, analyses will also be undertaken on the complete data which will include participants randomised after such an action was taken.

5 Statistical Modelling

In this section we outline considerations relevant to the statistical modelling approach used in M-FIT. These include a discussion on the model specifications, subgroup, secondary and other analyses. The economic analyses are not discussed in this SAP.

5.1 Primary Outcome Model

The primary analysis will use a longitudinal model for participant fatigue as measured by FACIT-Fatigue over the first 12 weeks of follow-up. All adaptations will be based on the relative intervention effects at 12 weeks post-randomisation, but we will also report the results at the other follow-up times.

While the FACIT-Fatigue score is ordinal in the range of 0-52, the primary analysis assumes a continuous response. This approach simplifies both the implementation and interpretation.

The final analysis will occur at the maximum sample size (or lower if early stopping occurs) after all participants have reached the primary endpoint and the follow-up data collected to 36 weeks. We will determine treatment comparative effectiveness based on the posterior distribution from the primary analysis. The interim analyses will use the same model specification as the final analysis; also see [Trial Adaptations and Statistical Decisions](#).

5.1.1 Model specification

The pre-specified primary model be a multivariate linear Normal model assuming an unstructured mean model for the outcomes at each visit occasion and treatment, and specifying an unstructured covariance matrix shared by all treatment groups.

Denote by $j \in \{0, 1, 2, 3, 4\}$ the visit occasion (0, 4, 8, 12, or 36 weeks), and by $k \in \{0, 1, 2, 3\}$ the treatment assigned to participant i and encoded by x_i . Denote by y_{ij} the FACIT-Fatigue score for participant i at visit j . Then, for $i = 1, \dots, n$ and $j = 0, 1, 2, 3, 4$,

$$\begin{aligned} \mu_{ij} | \alpha, \beta, \xi &= \alpha + (\beta_j + x_{k(i)}^\top \xi_j) \mathbf{1}_{j>0} \\ y_i | \mu_i, \Sigma &\sim \text{Normal}(\mu_i, \Sigma) \end{aligned}$$

where

$$\Sigma = \text{diag}(\sigma)\Omega\text{diag}(\sigma)$$

and $x_{k(i)}$ is the 1×3 intervention design vector (orthonormal) for participant i who was assigned to intervention $k(i)$. Similarly, let μ_{kj} be the expected FACIT-Fatigue score at visit j under treatment k .

The model parameters are:

- α - the expected FACIT-Fatigue at baseline (shared across all intervention groups)
- β_j - the change in FACIT-Fatigue from baseline to visit j on average (equally weighted) across all treatment groups
- ξ_j - the design coefficients for visit j
- Σ - residual covariance of FACIT-Fatigue.

A landmark analysis at visit $j = 3$ informs the primary comparison between treatments.

For priors, we specify

$$\begin{aligned}\sigma_j &\sim \text{Student-t}(3, 0, 5^2), \quad j = 1, 2, 3, 4 \\ \Omega &\sim \text{LKJ}(2) \\ \alpha &\sim \text{Normal}(\text{med}(y), \text{mad}(y)^2) \\ \beta_j &\sim \text{Normal}(0, 5^2), \quad j = 1, 2, 3, 4 \\ \xi_{pj} &\sim \text{Normal}(0, 5^2), \quad p = 1, 2, 3, \quad j = 1, 2, 3, 4\end{aligned}$$

where $\text{med}(y)$ and $\text{mad}(y)$ are the sample median and scaled median absolute deviation respectively.

Although the unstructured mean model is the proposed default, a more parsimonious specification may be appropriate in light of the observed data (e.g. linear or piece-wise linear with respect to visit). Any deviation from the primary model as specified should be justified (e.g. via model comparison metrics). Similarly, if issues are encountered with the unstructured covariance, then a more parsimonious structure might be assumed (e.g. AR(1))

5.2 Subgroup Analyses

Subgroup analyses for the primary model will examine heterogeneity in the FACIT-Fatigue score at 12 weeks post-randomisation arising from baseline exercise arm preference. Participants state their preference at baseline and are categorised according to:

1. no particular preference
2. preference for a particular exercise:
 - a) preference for walking
 - b) preference for resistance and aerobic
 - c) preference for resistance-only

Participants who have a preference may receive that intervention by chance when randomised. This defines an additional two groups:

1. participants who received their preference
2. participants did not receive their preference

A participant having a preference is a baseline variable (and say may shift the baseline FACIT-Fatigue score), but whether the participant receives their preferred exercise is only known post-randomisation (and will not shift baseline FACIT-Fatigue score).

A first subgroup analysis allows an expanded model where receiving the preferred exercise intervention may alter the effect of the intervention, but that any such effect is non-differential across the exercises. The expanded model is then

$$\mu_{ij} = (\alpha + \zeta \mathbf{h}\mathbf{p}_i) + (\beta_j + x_i^\top \xi_j + \zeta \mathbf{h}\mathbf{p}_i + \kappa_j \mathbf{r}\mathbf{p}_i) 1_{j>0}$$

where $\mathbf{h}\mathbf{p}_i \in \{0, 1\}$ indicates whether the participant expressed having a preference or not, and $\mathbf{r}\mathbf{p}_i \in \{0, 1\}$ indicates whether the participant received their preference. The model priors are analogous to the primary analysis model.

Alternatively, the actual intervention preferred may matter and the effect of receiving the preferred exercise may be allowed to vary according to the specific exercise by introducing an interaction term.

5.3 Secondary Analyses of Primary Outcome

Secondary analysis of the primary outcome will report on item specific summaries (i.e. mean and standard deviation of item specific scores by intervention and visit). Including ordinal mixed modelling of item specific scores accounting for within participant and within item clustering.

5.4 Secondary Outcome Analyses

The secondary outcomes are a mix of continuous, discrete and categorical variables, some collected at multiple timepoints over the duration of the study. In general, the analysis of the secondary outcomes will use mixed models which will be specified to adjust for baseline outcome, study arm, measurement occasion, and their interaction with measurement occasion, and correlation between measures on the same participant. For binary outcomes these regression models will be logistic and for ordinal outcomes cumulative logistic assuming proportional odds. Continuous outcomes may be analysed analogously as the primary outcome, or by linear mixed model.

The following provides brief specifications for the analyses of each of the secondary outcomes. The secondary outcomes are divided into groups consistent with their previous introduction in [3.2](#):

- [5.4.1 Patient-reported Outcomes](#)
- [5.4.3 Clinical Outcomes](#)
- [5.4.2 Exercise professional assessed outcomes](#)
- [3.2.6 Economic outcomes](#)

5.4.1 Patient-reported Outcomes

SONG-HD-F

There are two aims for the analysis on SONG-HD Fatigue:

1. compare fatigue at 4, 8, 12 and 36 weeks by intervention.
2. validation of the SONG-HD Fatigue score (WHAT METHODS ARE PROPOSED FOR THIS?)

We will analyse the SONG-HD-F score using an ordinal mixed effects model. The model will include baseline score as part of the response, and have fixed terms for visit, the interaction between intervention and follow-up visit (forcing the baseline distribution to be shared across all treatment groups), and random intercepts for participant and site. Data from each available measurement occasion will be included for each participant.

If issues are encountered, the model may be simplified by ignoring the site specific effects or by analysing visit differences independently (i.e. ignoring the longitudinal measurements e.g. comparison of interventions at 12 weeks ignoring the other repeated measurements).

PFS

The aim is to use PFS to compare fatigue at 4, 8, 12 and 36 weeks by intervention. Although the scores are ordinal (0 to 50), the Physical and mental fatigue scores will be analysed jointly as continuous outcomes using a multivariate linear model. The model will include fixed terms for arm, visit, and their interaction for the mean response. The model will assume unstructured covariance for responses within each participant. The outcomes at baseline will be constrained to have equal mean across all treatment groups. Data from all available measurement occasions will be included.

If issues are encountered with the model, then a more parsimonious correlation between outcomes may be assumed (e.g. Kronecker product of outcome and visit correlations, or assume independence between the outcomes, i.e. analyse them separately).

EQ-5D-5L

The aim is to use the EQ-5D-5L score to compare baseline adjusted quality of life at 4, 8, 12 and 36 weeks by intervention.

The amount of missingness and patterns of missingness will be reported, including, profile-item missingness and complete non-response. In analysing participant profiles, we may treat the entire profile as missing if any specific item is missing. Profiles will be reported by measurement occasion and arm, and in aggregate for each visit across all arms. These will be reported as frequencies and proportions of participants within each health-state ordered by frequency in aggregate, possibly restricted to the most frequently observed health-states.

Inference for intervention effect will be assessed by ordinal regression with fixed terms for visit, visit by intervention interaction, and EQ-5D-5L item, as well as random

intercepts for participant. The frequency of VAS responses will be presented for each arm and visit occasion. For inference, VAS scores will be analysed using linear mixed model with fixed terms for arm, visit, their interaction, and random intercepts and coefficients for visit by participant.

HADS

The aim is to use HADS to compare mood at 4, 8, 12 and 36 weeks by study arm. Anxiety and depression scores will be analysed as separate but correlated responses. Each response will be analysed using ordinal regression (cumulative logistic with proportional odds) with fixed terms for visit, visit by intervention interaction (with no main effect for intervention forcing shared baseline across randomised groups), participant specific intercepts which are correlated across the outcomes, and site specific intercepts.

If issues are encountered, the model may be simplified by ignoring the site specific effects or by analysing visit differences independently (i.e. ignoring the longitudinal measurements e.g. comparison of interventions at 12 weeks ignoring the other repeated measurements).

PROMIS-APSRA

The aim is to use the PROMIS Scale to compare social participation at 4, 8, 12 and 36 weeks by intervention. The aggregated score will be analysed using ordinal regression (cumulative logistic with proportional odds) with fixed terms for visit, visit by intervention interaction (with no main effect for intervention forcing shared baseline across randomised groups), and participant and site specific intercepts.

If issues are encountered, the model may be simplified by ignoring the site specific effects or by analysing visit differences independently (i.e. ignoring the longitudinal measurements e.g. comparison of interventions at 12 weeks ignoring the other repeated measurements).

mSUS

The items comprising mSUS will each be summarised in aggregate across the interventions.

5.4.2 Exercise professional assessed outcomes

Strength

The aim is to use several supervised tests to compare strength metrics at 12 and 36 weeks by intervention. The tests are:

- 30 second sit to stand (count)
- modified wall push-up test (count)
- timed up and go (seconds)
- arm curl test (count)
- hand grip strength test (kg)

The analyses...

Cardiorespiratory

The aim is to use the six-minute walk test (metres) to compare cardiorespiratory fitness at 12 and 36 weeks by intervention.

We will analyse distance covered during the six-minute walk test using log-normal mixed effect model for distance covered with fixed terms for treatment, visit, random intercept for participant repeat measures and random intercept for site.

Balance

The aim is to use a supervised Tinetti balance test to compare balance at 12 and 36 weeks by intervention. The Tinetti Test (also known as Performance Oriented Mobility Assessment) comprises two sections, one examining static balance abilities in a chair and then standing, and the other gait. Scoring is done on a three point ordinal scale with a range of 0 to 2. The maximum total score is 28 points (12 for gait, 16 for balance; higher scores are better).

The aggregated score will be analysed using ordinal regression (cumulative logistic with proportional odds) with fixed terms for visit, visit by intervention interaction (with no main effect for intervention forcing shared baseline across randomised groups), and participant and site specific intercepts.

If issues are encountered, the model may be simplified by ignoring the site specific effects or by analysing visit differences independently (i.e. ignoring the longitudinal

measurements e.g. comparison of interventions at 12 weeks ignoring the other repeated measurements).

Frailty

The aim is to use the Fried Frailty Index to compare patient baseline adjusted frailty at 12 and 36 weeks by intervention. The Fried Frailty Index comprises five criteria for assessing weight loss, exhaustion, physical activity, slowness and weakness. The sum of scores in the categories classifies individuals into three frailty conditions: not frail, pre-frail and frail.

We will analyse FFI using an ordinal mixed effects models for frailty with fixed terms for treatment, timepoint, random intercept for participant repeat measures and random intercept for exercise physiologist.

Body composition

The aim is to use the BMI and waist circumference to compare body composition at 12 and 36 weeks by intervention. In order to compute body composition metrics, height (cm), weight (kg) and waist circumference (cm) will be collected.

We will analyse BMI and waist circumference using methods analogous to the primary outcome, i.e. using multivariate linear models.

5.4.3 Clinical Outcomes

Standardised Outcomes in Nephrology (SONG) core outcomes follow.

Vascular access function

The aim is to compare the rate of vascular access repairs at 36 weeks by intervention.

We will analyse the frequency of repairs using Poisson (Negative Binomial if overdispersion is apparent) regression with fixed terms for intervention and random intercept for site.

Technique survival

TODO

Peritoneal dialysis infections

The aim is to compare the rates of PD associated infection at 36 weeks by intervention. Typical rates of PD associated infection are around 0.24-1.66 episodes/patient/year.

We will analyse the frequency of infections using Poisson (Negative Binomial if overdispersion is apparent) regression with fixed terms for intervention and random intercept for site.

Physical activity

The aim is to use Actigraph/Fitbit data to compare the duration of moderate and vigorous physical at 12 and 36 weeks by intervention. These analyses will be exploratory using the actigraph and fitbit data.

Sleep

The aim is to compare duration of sleep (minutes) at 12 and 36 weeks by intervention. These analyses will be exploratory using the actigraph and fitbit data.

5.5 Model Checks

The assumed models will be assessed for adequacy. Outliers and influential data points will be investigated. Additional models (either simpler or more complex) may be investigated as part of checks of sensitivity, stability, and model fit. If any issues or concerns arise, all changes or updates to the specified primary model will be documented and reported including justification of the changes. Model assessment and comparison will be via posterior predictive checks and cross-validation.

For the primary outcome, if there are issues with convergence or stability of the primary model, then the primary analysis may revert to more parsimonious model (e.g. altered covariance structure, or simplified mean specification).

For the secondary outcomes, models for which proportional odds is assumed will have the assumption assessed. If the assumption appears invalid than partial proportional odds models may be used instead, or outcomes at each time point may be analysed separately.

Any deviations from the pre-specified model will be reported along with the reasons for deviating and why the alternative model is preferred.

5.6 Qualitative Analyses

Evaluation of the impact, fidelity, facilitators, and barriers of implementing the exercise program in patients receiving dialysis. These analyses are beyond the scope of the SAP.

5.7 Missing Data

Dialysis patients are generally not lost to follow up due to the nature of their treatment protocols. Withdrawal of consent is also expected to be low in this population. However, adherence can be a problem and safety is an ongoing concern. Interim missingness might also occur if participants are on occasion unable or unwilling to complete the study questionnaires within the protocol timeframe. Such interim missingness could be differential between treatments groups. For example, if a particular exercise regimen does reduce fatigue, participants may be more willing to or capable of completing the requisite survey items compared to participants who are experiencing greater fatigue under a different intervention. This would which correspond with outcomes missing not at random (MNAR).

The primary analysis will assume that outcomes are missing at random (MAR) given the specified covariates. Sensitivity analyses will investigate the sensitivity of inferences under various MNAR assumptions using multiple imputation. This will include δ -adjustment methods to assess robustness of effectiveness results whereby observations missing due to loss-to-follow-up will be assumed to have reduced FACIT-Fatigue scores by an amount δ relative to the assumed MAR response [23].

5.7.1 Baseline

Patterns of missingness for the baseline data will be reported by intervention group. Baseline covariates which are to be included in the pre-specified models will be imputed, if missing, using information in the other baseline covariates. Study outcomes measured at baseline will be included in the response model.

5.7.2 Follow-up

Patterns of missingness for follow-up data will be reported and summarised by intervention group across all measurement occasions. We will investigate the relationships between baseline covariates and drop-out or interim missingness of follow-up observations.

N (%)	Baseline	Week 4	Week 8	Week 12	Week 36
xx (xx)	✓	✓	✓	✓	✓
xx (xx)	✓	✓	✓		
xx (xx)	✓	✓			
xx (xx)	✓				
xx (xx)		✓	✓	✓	
xx (xx)	✓		✓		
		etc.			

Table 2: Example summary of data missingness for study outcome. Observed values indicated by ✓.

5.8 Sensitivity Analyses

Sensitivity of treatment effect estimates to the prior on the treatment parameters will be assessed via importance sampling or independent model fits where necessary. The prior variance for the treatment specific terms will be varied and the posterior summaries reported as a function of this variance.

Sensitivity to the MAR assumption will be investigated by allowing MNAR via pattern mixture modelling using reference-based analyses.

For example, in the intent-to-treat (ITT) analysis, participants who drop-out may be assumed to have ceased exercise, and their expected FACIT-Fatigue scores to have reverted to the baseline distribution. Alternatively, participants on active treatment who drop-out may be assumed to have similar outcomes to those on control. If there are participants with follow-up who are known to have not adhered to their assigned intervention, another option may be to use this data to inform expected responses in participants post drop-out. However, this will be conditional on the availability of such data. Methods of imputation will be reported were utilised.

5.9 Safety Analyses

The following safety data is of interest and recorded at 4, 8, 12 and 36 weeks (the Pre-exercise safety questionnaire is also collected at baseline).

- Death
- Cardiac events
- Falls
- Exercise related injuries
- Pre-exercise safety questionnaire

Descriptive statistics of each safety variable will be tabulated in aggregate and by group and follow-up. The frequency and proportion of participants with any adverse event, and the frequency and rate of adverse events will be reported.

5.10 Exploratory Analyses

For longitudinal ordinal outcomes, we will investigate analyses via ordinal transition models as a comparator for the use of ordinal mixed effects models.

6 Statistical Quantities

Statistical quantities obtained from the primary model will be used to evaluate treatment effectiveness at interim analyses and direct the progression and adaptations for the trial. In what follows, \mathcal{A}_{t-1} is used to denote the available exercise interventions following the $t - 1$ th interim analysis occasion with $\mathcal{A}_0 = \{1, 2, 3\}$ at trial initiation. Note that this set excludes the attention-control.

6.1 Comparison with best and control

Section 5.1 introduced μ_{kj} as the mean FACIT-Fatigue score under intervention k at visit j . The probability that intervention k is best is defined according to

$$\begin{aligned}\delta_{kt} &= \mu_{k3} - \max_{k' \in \mathcal{A}_{t-1} \setminus \{k\}} \mu_{k'3} \\ \pi_{kt} &= \Pr[\delta_{kt} > 0 | \text{data}_t], \quad k \in \mathcal{A}_{t-1}\end{aligned}$$

which will informally be referred to as the probability of being best amongst the active arms. For $k \notin \mathcal{A}_t$ we set $\pi_{kt} = 0$.

Decision rules are also defined in terms of comparisons with attention-control. The probability that intervention k is effective (has higher FACIT-Fatigue scores than attention control) is defined by

$$\omega_{kt} = \Pr[\mu_{kt} - \mu_{0t} > 0 | \text{data}_t], \quad k = 1, 2, 3.$$

6.2 Intervention superiority

A decision of intervention superiority for intervention k relative to all other interventions is recommended when intervention k has high probability of being best. Specifically, the superiority decision rule is defined by

$$\pi_{kt} > \varepsilon_{\text{sup}} = 0.98$$

where $\varepsilon_{\text{sup}} = 0.98$ is the superiority decision threshold, selected through trial simulation.

6.3 Intervention inferiority

A decision of intervention inferiority for intervention k relative to all other interventions is recommended when intervention k has low probability of being best. Specifically, the inferiority decision rule is defined by

$$\pi_{kt} < (1 - \varepsilon_{\text{sup}})/(|\mathcal{A}_t| - 1) = 0.02/(|\mathcal{A}_t| - 1).$$

Under this decision threshold, when all interventions bar one are inferior, the remaining intervention satisfies superiority.

6.4 Intervention effectiveness

The decision rule for effectiveness of an intervention (relative to attention-control) is defined by

$$\omega_{kt} > \varepsilon_{\text{eff}} = 0.98$$

in which case a decision of intervention effectiveness is recommended.

6.5 Intervention ineffectiveness

The decision rule for ineffectiveness of an intervention (relative to attention-control) is defined by

$$\omega_{kt} < 1 - \varepsilon_{\text{eff}} = 0.02$$

in which case, a decision of intervention ineffectiveness is recommended.

7 Trial Adaptations and Statistical Decisions

As the trial proceeds, the accrued information is used to make decisions on the progression of the trial based on pre-specified adaptations and decision rules.

For adaptations internal to the trial, predefined rules are in place to inform trial decisions conditional on the primary model. Pending independent oversight from a DSMC, these statistical decisions will inform conclusions such as deciding arms to be superior or inferior. The following sections outline these adaptations.

7.1 Sequential Analyses

Regular interim analyses will be scheduled according to the sample size of participants with follow-up to the primary endpoint. The first analysis will occur when $n = 100$ participants have reached the 12 week primary endpoint (regardless of whether the outcome was observed or not). After the first analysis, interim analyses will occur when $n = 200$ and $n = 300$ participants have reached the primary endpoint. The final analysis will occur when follow-up has been completed on participants after terminating recruitment due to either reaching the maximum sample size of 400 or stopping earlier.

The analyses will include the data from all the enrolled participants at the time of the analysis. Some of the participants will only contribute their baseline data, some will contribute their baseline and 4 week follow up and so on up to their 12 week outcome.

7.2 Stopping rules

Enrollment into the trial may be stopped prior to the maximum sample size if:

- one of the exercise arms is declared superior and effective
- all exercise arms have been declared inferior or ineffective

The trial may be stopped early due to safety concerns or other reasons at the discretion of the relevant decision making body. Otherwise the trial will continue until the maximum sample size.

7.3 Response-Adaptive Randomisation

Target allocation to exercise arms will be updated to be proportional to a transformation of the probability an exercise arm is the best (in terms of reducing fatigue), the variance of the posterior change in fatigue divided, and the sample size on that arm. Note that the response adaptive randomisation is based on the comparison amongst the active exercise arms at the week 12 follow up.

Following an interim analysis, an exercise arm may be dropped if:

- it is inferior to the other exercise arms (probability of it being superior below decision threshold)
- it is inferior to the attention-control arm (probability of it being effective below decision threshold).

Denote by \mathcal{A}_t the set of exercise arm indices which will continue in the trial following the current interim t . If an exercise arm satisfies either of the above criteria, then it may be removed from \mathcal{A}_t .

We will compute allocation probabilities according to

$$r_{t0} = \frac{1}{|\mathcal{A}_t + 1|}$$

$$r_{tk} = \begin{cases} 0 & \text{if } k \notin \mathcal{A}_t \\ (1 - r_0) \frac{f(\pi_k, n_k)}{\sum_{j=1}^3 f(\pi_j, n_j)} & \text{if } k \in \mathcal{A}_t \end{cases}, \quad k = 1, 2, 3.$$

where $f(\pi, n) = \sqrt{\pi / \max(n, 1)}$, and $\mathcal{A}_0 = \{1, 2, 3\}$ denotes the available exercise arms at trial start. Once an arm is dropped from the trial, it is unavailable for the remainder of the trial, that is, $\mathcal{A}_{t+1} \subseteq \mathcal{A}_t$.

In practice, the target allocations may be updated prior to confirmation of the decision on arm dropping (e.g. TSC have final decision in whether to drop an arm), in which case the relevant values of r_t will be set to zero only after such a confirmation has been made, and the remaining arm allocations re-normalised to sum-to-one.

In essence, two updates would then occur. The first being an update without any arm

dropping

$$r_{t0}^* = r_{t-1,0}$$

$$r_{tk}^* = (1 - r_{t0}^*) \frac{f(\pi_k, n_k)}{\sum_{k=1}^3 f(\pi_k, n_k)}, \quad k \in \mathcal{A}_{t-1}$$

and then following confirmation that arms are to be dropped

$$r_{t0} = \frac{1}{|\mathcal{A}_t + 1|}$$

$$r_{tk} = \begin{cases} 0 & \text{if } k \notin \mathcal{A}_t \\ (1 - r_{t0}) \frac{r_{tk}^*}{\sum_{j \in \mathcal{A}_t} r_{tj}^*} & \text{if } k \in \mathcal{A}_t. \end{cases}, \quad k = 1, 2, 3.$$

In this manner, if a decision is made to *not* drop an arm despite it meeting the criteria, the update to allocations would not need to be revised and there is no delay in applying the updated allocation ratios whilst confirmation of arm dropping is pending.

8 Interim Analyses and Trial Reporting

Following each scheduled interim analysis, a closed DSMC report will be generated with the following content:

- Executive summary
- Protocol synopsis
- Report overview
- Suggested communication to the study team investigators
- Enrollment status including rates of enrollment
- Participant status
- Data completeness
- Descriptive statistics for baseline demographic, co-morbidities, and other baseline factors, stratified by treatment arm and in aggregate
- Descriptive statistics of adherence to treatment protocol, protocol violations and dropouts
- Safety data
- Protocol deviations
- Descriptive statistics pertaining to the primary outcome
- Inference results from primary analysis
- Appendix: As required

The detail of report content will be agreed upon with the DSMC prior to the start of recruitment.

8.1 Trial Adaptations

The DSMC's role and responsibility in recommending trial adaptations will be outlined in the DSMC charter.

8.2 Trial Reporting

Communication of trial outcomes will not occur until after termination of recruitment.

References

1. Yellen, S. B., Cella, D. F., Webster, K., Blendowski, C. & Kaplan, E. Measuring fatigue and other anemia-related symptoms with the Functional Assessment of Cancer Therapy (FACT) measurement system. *Journal of pain and symptom management* **13**, 63–74 (1997).
2. Cella, D., Lai, J.-s., Chang, C.-H., Peterman, A. & Slavin, M. Fatigue in cancer patients compared with fatigue in the general United States population. *Cancer* **94**, 528–538 (2002).
3. Montan, I., Löwe, B., Cella, D., Mehnert, A. & Hinz, A. General population norms for the functional assessment of chronic illness therapy (FACIT)-fatigue scale. *Value in Health* **21**, 1313–1321 (2018).
4. Wang, S.-Y. *et al.* Psychometric properties of the Functional Assessment of Chronic Illness Therapy-Fatigue (FACIT-Fatigue) in Chinese patients receiving maintenance dialysis. *Journal of pain and symptom management* **49**, 135–143 (2015).
5. Ju, A. *et al.* Establishing a core outcome measure for fatigue in patients on hemodialysis: a Standardized Outcomes in Nephrology–Hemodialysis (SONG-HD) consensus workshop report. *American Journal of Kidney Diseases* **72**, 104–112 (2018).
6. Glynn, N. W. *et al.* The Pittsburgh Fatigability scale for older adults: development and validation. *Journal of the American Geriatrics Society* **63**, 130–135 (2015).
7. Morton, F. & Nijjar, J. S. *eq5d: Methods for Analysing 'EQ-5D' Data and Calculating 'EQ-5D' Index Scores* R package version 0.10.1 (2021). <https://CRAN.R-project.org/package=eq5d>.
8. Devlin, N., Parkin, D. & Janssen, B. *Methods for analysing and reporting EQ-5D data* (Springer Nature, 2020).
9. Zigmond, A. S. & Snaith, R. P. The hospital anxiety and depression scale. *Acta psychiatrica scandinavica* **67**, 361–370 (1983).
10. Cella, D. *et al.* PROMIS[®] adult health profiles: efficient short-form measures of seven health domains. *Value in health* **22**, 537–544 (2019).
11. Rikli, R. E. & Jones, C. J. Development and validation of a functional fitness test for community-residing older adults. *Journal of aging and physical activity* **7**, 129–161 (1999).

12. Jones, C. J., Rikli, R. E. & Beam, W. C. A 30-s chair-stand test as a measure of lower body strength in community-residing older adults. *Research quarterly for exercise and sport* **70**, 113–119 (1999).
13. Macfarlane, D., Chou, K. L., Cheng, Y. & Chi, I. Validity and normative data for thirty-second chair stand test in elderly community-dwelling Hong Kong Chinese. *American Journal of Human Biology: The Official Journal of the Human Biology Association* **18**, 418–421 (2006).
14. Podsiadlo, D. & Richardson, S. The timed Up & Go: a test of basic functional mobility for frail elderly persons. *Journal of the American geriatrics Society* **39**, 142–148 (1991).
15. Shumway-Cook, A., Brauer, S. & Woollacott, M. Predicting the probability for falls in community-dwelling older adults using the Timed Up & Go Test. *Physical therapy* **80**, 896–903 (2000).
16. Liu, C.-j. *et al.* Predicting hand function in older adults: evaluations of grip strength, arm curl strength, and manual dexterity. *Aging clinical and experimental research* **29**, 753–760 (2017).
17. Butland, R., Pang, J., Gross, E., Woodcock, A. & Geddes, D. Two-, six-, and 12-minute walking tests in respiratory disease. *British medical journal (Clinical research ed.)* **284**, 1607 (1982).
18. Buar Pajek, M. *et al.* Six-minute walk test in renal failure patients: representative results, performance analysis and perceived dyspnea predictors. *PloS one* **11**, e0150414 (2016).
19. Tinetti, M. E., Williams, T. F. & Mayewski, R. Fall risk index for elderly patients based on number of chronic disabilities. *The American journal of medicine* **80**, 429–434 (1986).
20. Wyngaert, K. V. *et al.* Associations between the measures of physical function, risk of falls and the quality of life in haemodialysis patients: a cross-sectional study. *BMC nephrology* **21**, 1–10 (2020).
21. Fried, L. P. *et al.* Frailty in older adults: evidence for a phenotype. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* **56**, M146–M157 (2001).
22. Het Veld, L. P. O. *et al.* Fried phenotype of frailty: cross-sectional comparison of three frailty stages on various health domains. *BMC geriatrics* **15**, 1–11 (2015).
23. Carpenter, J. & Kenward, M. *Multiple imputation and its application* (John Wiley & Sons, 2012).

Appendix I: Trial Operating Characteristics

Simulation Model

The analysis model used in the simulations ignored the longitudinal outcomes,

$$\begin{aligned} y_{i,3} | \alpha, \beta, \sigma &\sim \text{Normal}(\alpha + x_{\text{trt}[i]}^\top \beta, \sigma^2) \\ \mu_{j,3} &= \alpha + x_j^\top \beta \\ \alpha &\sim \text{Normal}(40, 5) \\ \beta &\sim \text{Normal}(0, 5) \\ \sigma &\sim \text{Half-}t(3, 0, 5) \end{aligned}$$

where $x_{\text{trt}[i]}$ for $\text{trt}[i] \in \{0, 1, 2, 3\}$ denotes the treatment design vector corresponding to participants i 's assigned treatment. For simplicity, this model was also used for data generation, despite the true outcome being discrete and bounded between 0 to 52.

The primary quantity of interest is the difference in FACIT-Fatigue at 12-weeks for each active treatment group relative to the control group, that is,

$$\Delta_j = \mu_{j,3} - \mu_{0,3}, \quad j = 1, 2, 3.$$

Another quantity of interest is the relative effectiveness of each active treatment compared to each other active treatment.

$$\delta_j = \mu_{j,3} - \max_{j' \neq j} \mu_{j',3}, \quad j = 1, 2, 3.$$

We quantify their probability under the assumed model according to

$$\begin{aligned} \omega_j &= \Pr(\Delta_j > 0 | \text{data}) \\ \pi_j &= \Pr(\delta_j > 0 | \text{data}) \end{aligned}$$

such that ω_j is the posterior probability that treatment j is effective, and π_j is the posterior probability that treatment j is best amongst the active treatments.

Interim Analyses

Interim analyses are scheduled to occur when 100, 200, 300 and 400 participants reach their primary endpoint. At each interim analysis, we pre-specify a number of decision rules in terms of the previously defined quantities. These decision rules may be to drop specific treatment arms, or to stop the trial altogether. We denote by J' the number of available arms at the time of the interim analysis, including control. At the first interim analysis, $J' = 4$.

- *Effective*: if $\omega_j > \epsilon_0$, then treatment j has probability greater than ϵ_0 of being effective compared to control.
- *Ineffective*: if $\omega_j < 1 - \epsilon_1$ then treatment j has probability greater than ϵ_1 of being ineffective compared to control, and therefore may be dropped from the trial.
- *Superiority*: if $\pi_j > \epsilon_2$, then treatment j has probability greater than ϵ_1 of being the best active treatment, and therefore, all other active treatments may be dropped from the trial.
- *Inferiority*: if $\pi_j < \epsilon_2/(J' - 2)$, then treatment j has probability less than $\epsilon_1/(J' - 2)$ of being the best active treatment, and therefore may be dropped from the trial.
- *Stopping*: the trial may be stopped early if one active treatment is superior and effective.

Response-Adaptive Randomisation

Initially, each treatment arm are allocated with equal-probability of $r_j = 1/J'$ where $J' = 4$. The probability of allocation to the control group is fixed throughout the trial to be equal to the reciprocal of the number of arms still on study, $r_0 = 1/J'$.

For the remaining arms, the allocation may be zero if the arm has been dropped, or it may be proportional to the probability it is the best active treatment.

Following each interim analysis, the allocation probability of the 3 active arms are updated according to

$$r_j = \begin{cases} 0 & \text{if the arm has been dropped} \\ \frac{(\pi_j/n_j)^k}{\sum_{j=1}^3 (\pi_j/n_j)^k} & \text{otherwise} \end{cases}, \quad j = 1, 2, 3.$$

for some scaling factor k . If $k = 0$, then $r_j = 1/J'$ for each arm and if $k = 1$ then $r_j = \pi_j$. The simulations used $k = 0.5$.

Simulations

The simulations presented here assume:

- expected accrual of 3 participants per week according to homogeneous Poisson process.
- interim analysis when 100, 200, 300, and 400 participants reach their primary endpoint.
- treatment effect of 0, 0.5, or 1 standard deviation on 1, 2, or 3 arms.
- 20% drop-out, so that, for example when 100 participants reach their primary endpoint, 20 have missing 12-week FACIT-Fatigue data.
- decision thresholds of $\epsilon_0 = \epsilon_1 = \epsilon_2 = 0.98$.
- ignores repeated measures analysing 12-week FACIT-Fatigue as response in a linear model conditional on treatment group
- uses variational approximation for model posterior.

Table 3: Summary of trial operating characteristics, n = 10,000 simulations, $\epsilon_0 = \epsilon_1 = \epsilon_2 = 0.98$.

Effect	Number	Treatment	Allocated	Superior	Inferior	Effective	Ineffective	Active
0	0	0 - control	83					
		1	79	0.01	0.10	0.02	0.05	0.90
		2	78	0.01	0.10	0.02	0.05	0.89
		3	79	0.01	0.10	0.02	0.05	0.89
0.5SD	1	0 - control	71					
		1	93	0.83	0.00	0.92	0.00	1.00
		2	46	0.00	0.89	0.01	0.04	0.11
		3	46	0.00	0.89	0.01	0.04	0.11
	2	0 - control	89					
		1	92	0.05	0.05	0.86	0.00	0.95
		2	93	0.05	0.05	0.87	0.00	0.95
		3	37	0.00	0.94	0.01	0.04	0.06
	3	0 - control	82					
		1	78	0.01	0.08	0.79	0.00	0.92
		2	78	0.01	0.08	0.78	0.00	0.92
		3	79	0.01	0.08	0.80	0.00	0.92
1SD	1	0 - control	37					
		1	42	1.00	0.00	1.00	0.00	1.00
		2	31	0.00	1.00	0.01	0.03	0.00
		3	31	0.00	1.00	0.01	0.03	0.00
	2	0 - control	92					
		1	93	0.06	0.06	0.99	0.00	0.94
		2	92	0.06	0.06	1.00	0.00	0.94
		3	30	0.00	1.00	0.02	0.05	0.00
	3	0 - control	81					
		1	78	0.01	0.09	1.00	0.00	0.91
		2	79	0.01	0.09	1.00	0.00	0.91
		3	79	0.01	0.09	1.00	0.00	0.91

Appendix II: Study Database and Derived Variables

The following sections summarise the relationship between the database variables and the variables outlined in the protocol and SAP. Where required, the formula for calculating the outcome using the database labels is included.

Outcomes

FACIT-Fatigue

(NOTE AS OF WRITING, THE DATABASE IS MISSING 2 ITEMS FOR FACIT-FATIGUE)

Responses to the FACIT-Fatigue survey are recorded as scores of 1-5 in the database. Therefore, to calculate the score, 1 must be subtracted from each item. Some items also require value reversals. If an item is missing, the item is excluded and the denominator is decreased by 1. If fewer than 50% (i.e. 6 or fewer) of items were answered, then the score is set to missing.

```
responses_ff = # of completed items  
sum_ff = ... + fatigue_ff + weak_ff + listless_ff + tired_ff  
          + trouble_start_ff + trouble_finis_ff - energy_ff +  
          - able_to_do_ff + help_ff + frustrated_ff + social_ff
```

```
score_ff = if (response_ff > 6) 13*sum_ff/response_ff else NA
```