STATISTICAL ANALYSIS PLAN (SAP)

Protocol Title	Statistical analysis plan for the STEPWISE Study: a Smartphone-Based Exercise Solution for People with Parkinson's Disease (randomized controlled trial)
Protocol Number	NL75501.091.20
Protocol Version	11.0, 29 Nov 2023
Trial registration	NCT04848077
Regulatory Sponsor	Bastiaan R. Bloem, MD, PhD
Supported by	The Netherlands Organisation for Health Research and Development; grant number 546003007
SAP Version	1.0, 08 August 2024

SAP APPROVAL SIGNATURES

Bastiaan R. Bloem, MD, PhD
Principal investigator
Date 13-09-2024

Date: 08-08-2024

Nienke M. de Vries, PhD
Coordinating investigator

Kit B. Roes, PhD / Eric A. Macklin, PhD
Biostatistician
Date 04-Sep-2024

Sabine Schootemeijer, MSc.
PhD student
Date 16-09-2024

SAP REVISION HISTORY

Version	Date	Description of Changes
1.0	12 Dec 2023	Initial draft
	5 Feb 2024	Changes after input Eric Macklin and Kit Roes
	8 Feb 2024	Include protocol amendment blood biomarkers
	26 March 2024	Include decision on covariates
	22 April 2024	Added medication dosage according to Jost et al 2023
	29 May 2024	Implemented decision on primary secondary outcome (MDS-UPDRS IB+II)
	07 August 2024	Implemented decision on adjustment MDS-UPDRS-III for LEDD

Table of Contents

1. Introduction

This statistical analysis plan (SAP) defines the outcome measures and analysis samples and specifies the planned analyses of data for the STEPWISE trial. The SAP supplements the clinical protocol (Schootemeijer et al., 2023). In case of discrepancies between the SAP and the clinical protocol concerning matters of data analysis, the SAP is authoritative. On all other matters, the clinical protocol is authoritative. This SAP specifies data and planned analyses for the main trial. Specification of data and analyses for ancillary studies will be detailed in ancillary SAPs if not covered here.

1.1. Background and rationale

Exercise has various health benefits for people with Parkinson's disease (PD). However, implementing exercise into daily life and long-term adherence remain challenging. To increase a sustainable engagement with physical activity of people with PD, interventions that are motivating, accessible, and scalable are needed. Recent innovations in digital technology, such as apps and sensors on smartwatches and smartphones, open up exciting avenues for remote interventions as well as remote monitoring of the outcome. So far, only one study has investigated the effectiveness of a (tablet-based) application in promoting physical activity in PD (Ellis et al., 2019). While this study showed that people with PD were satisfied with using an exercise app, a statistically insignificant change in physical activity was reported. Other studies in older adults showed that apps increased physical activity, but the interventions were of short duration (lasted for two to six months) (Yerrakalva et al., 2019). So, even though innovative technologies are highly promising, changing physical activity behavior in the long term is still a major challenge and needs further study.

Please refer to the trial protocol for more details on the rationale for the STEPWISE intervention (Schootemeijer et al., 2023).

1.2. Objectives

In this trial, we investigate the feasibility of a smartphone app (Smartphone-Titrated Exercise in Parkinson's With Incentive-Supported Engagement: STEPWISE app) to improve physical activity in people with PD. The primary objective is to evaluate the between-group difference in average daily step count change from baseline to one year post randomization (52 weeks) in participants assigned to any of three different intervention arms or a control group. The secondary objective is to investigate the effect of assignment to any of three different intervention arms or a control group

on change in secondary outcome measures (measures of physical fitness, PD symptoms, health-related quality of life, balance, gait speed, handgrip strength, falls, fear of falling, belief in ones' physical capacities, apathy, autonomic dysfunction, sleep, cognition, anxiety, depression, fatigue and blood-based biomarkers) from baseline to follow-up (~52 weeks after randomization). Our third objective is to explore whether there is a dose-response relationship between change in step counts from baseline to one year post randomization and our secondary outcome measures. Our primary safety objective is to investigate whether participants assigned to any of three different intervention arms experience more falls, or other adverse events, than the control group during the intervention period (52 weeks). Supportive analyses will further investigate temporal changes in step counts, associations between potential temporal changes in step counts and secondary outcome measures, and associations between step count volume (rather than change in step count from baseline) and secondary outcome measures.

1.3. Aims

Our primary aim is to evaluate whether dose-dependent encouragement through the STEPWISE app yields a sustained increase in step count, measured as the 52-week change in step counts. The primary estimate addressing this question will be the comparison of 52-week change in step counts between the very large increase and active control groups. The primary aim is further addressed by comparing the large increase group versus the active control group and the moderate increase group versus the active control group. The remaining comparisons are described as supportive step count aims (below).

Our secondary aim is to evaluate whether dose-dependent encouragement through the STEPWISE app yields an effect on 52-week change in secondary outcome measures of physical fitness (VO2max, six minute walk test), PD symptoms (Movement Disorders Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS), health-related quality of life (Parkinson's Disease Questionnaire-39; PDQ-39), balance (MiniBESTest), gait speed (10 meter walk test), handgrip strength, falls, fear of falling (Falls Efficacy Scale-International; FES-I), belief in ones' physical capacities (Lichamelijke Vaardigheden Schaal; LIVAS), apathy (Abbreviated version of the Apathy Evaluation Scale; AES12-PD), autonomic dysfunction and sleep (Scales for Outcomes in Parkinson's Disease; SCOPA), cognition (Montreal Cognitive Assessment; MoCA), anxiety and depression (Hamilton Anxiety and Depression Scale; HADS), fatigue (Fatigue Severity Scale; FSS) or blood-based biomarkers (panels to be determined). The primary estimate for intermediary

effects will be the comparison of change in six minute walking distance between the interventional arms combined (moderate, large, and very large increase) and the active control group. The primary estimate for efficacy will be the comparison of motor- and non-motor aspects of experiences of daily living (sum of MDS-UPDRS parts IB and II) between the interventional arms combined (moderate, large and very large increase) and the active control group.

Our third aim is to evaluate whether there is a dose-response relationship between absolute 52-week step count change and secondary outcomes (described for our secondary aim). The primary estimate for intermediary dose-response relationships will be the 52-week change in six minute walking distance for a given absolute 52-week change in step count, adjusted for age, sex, disease duration, and baseline step count. The primary estimate for dose-response efficacy will be the 52-week change in motor- and non-motor aspects of experiences of daily living (sum of MDS-UPDRS parts IB and II) for a given 52-week change in step count, adjusted for age, sex, disease duration, and baseline step count. Although we previously indicated (Schootemeijer et al., 2023) that we would adjust for baseline VO_{2max} when estimating the intermediary dose-response effects and dose-response efficacy, we will not adjust for VO_{2max} since only a subsample of 100 participants performed the VO_{2max}. In a supportive analysis, we will adjust for baseline VO_{2max} in this subsample.

For our primary safety aim, we will investigate whether participants assigned to any of three different intervention arms experience more falls, or other adverse events, than the active control group during the study period (53 weeks).

In addition, we propose several supportive analyses. Supportive analyses for step count (primary aim) include investigating the step count changes over different time courses (e.g., baseline to 16-week follow-up, baseline to 24-week follow-up), the (change in) distribution of step counts within weeks (increase in step counts on weekdays vs weekend days), and the relationships between step count changes and employment, living situation, sex, and minutes spent in total self-reported physical and sports activities.

Supportive analyses for our third aim include relating the volume increase in step count at different time points (e.g. baseline to 16-week follow-up, baseline to 24-week follow-up) with secondary outcome measures and relating the cumulative step count volume (rather than volume *increase*) with secondary outcome measures.

2. Study Methods

2.1. Trial design

STEPWISE is a double-blind, parallel-group, randomized controlled trial in people with PD who perform a limited volume of physical activities at baseline. The intervention consists of a motivational app (STEPWISE app) that aims to motivate people with PD to walk more. The STEPWISE app (Figure 1) contains several motivational elements to increase engagement. Participants receive feedback on the achieved percentage of their weekly step count target. Their step count target is visualized as a percentage of steps taken towards their step count target every week. Participants also see the number of steps they took that day, the day before and the cumulative steps during the study. Participants are encouraged to reach 100% of their step count target every week. They see their progression as a percentage of their weekly target rather than as an absolute step count or as a percentage of their baseline step count in order to blind them as much as possible.

Participants are instructed not to participate in other interventional studies for the duration of the study period (one year). They receive their regular care, which may consist of medication changes, outside of the study. These changes are reported at the follow-up visit.

The study is performed at Radboud University Medical Center (Radboudumc) and Canisius Wilhelmina Ziekenhuis (CWZ), Nijmegen, the Netherlands. Radboudumc is the study sponsor and is responsible for recruitment and inclusion of participants. Cardiorespiratory fitness (VO_{2max}) will be assessed among a subset of 100 participants at CWZ.

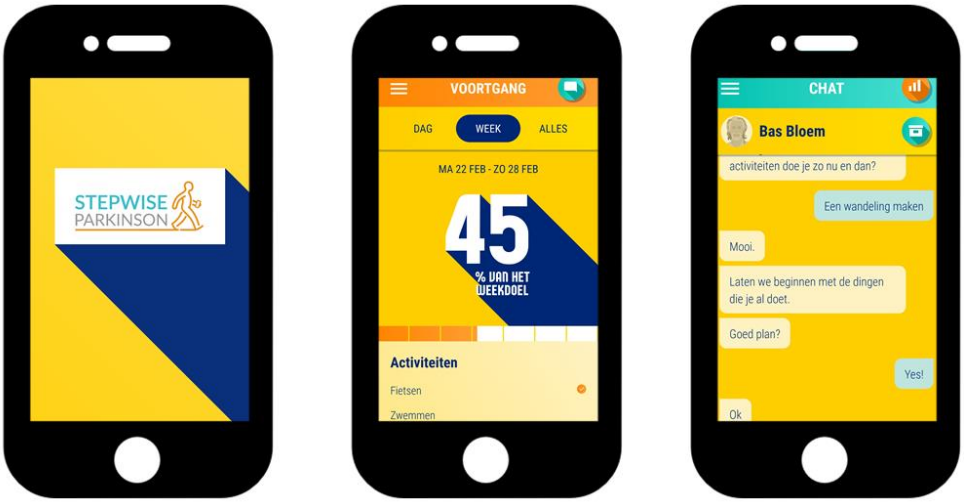

Figure 1. Screenshots STEPWISE Parkinson app. Splash screen, progression towards target, and chat with virtual coach.

2.2. Randomization

If participants are determined eligible (see eligibility criteria, §4.2), they complete a baseline set of assessments at Radboudumc and one week later, they are randomized to one of four treatment groups in a 1:1:1:1 ratio: the active control group (a small increase relative to their own step count at baseline) or to one of three intervention groups, each with a different step count increase (a moderate, a large, or a very large increase relative to their own step count at baseline; Figure 2 and 3) using the CastorEDC data management system (Castor, 2019). The randomization schedule uses random permuted blocks (block sizes: 4, 8, 12) stratified by sex (two groups: female and male) and disease duration (three groups: <5 years, 5-10 years, and >10 years disease duration).

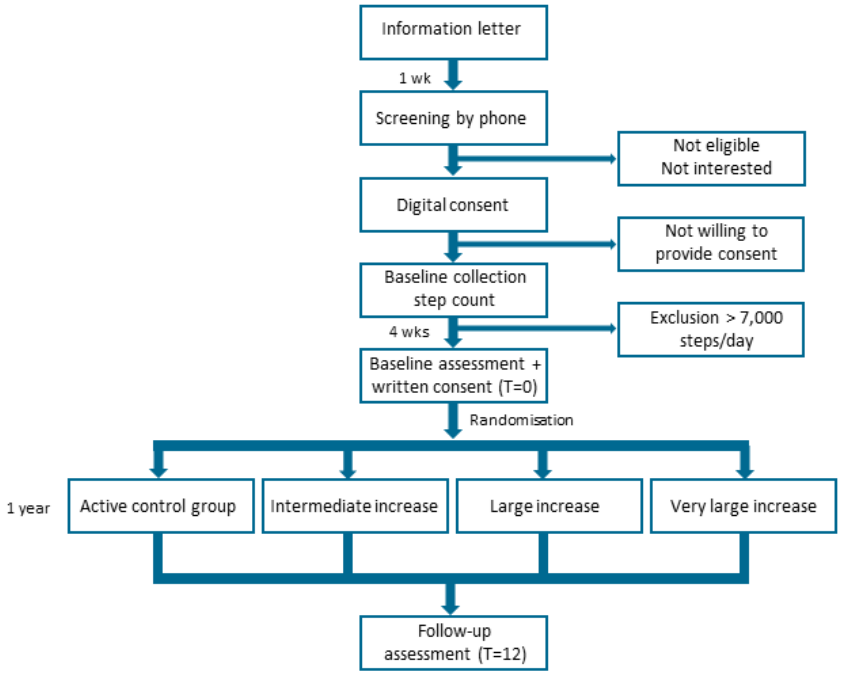

Figure 2. Participant flow chart.

2.3. Description of the intervention groups and control group

The targeted daily step count will be determined by incrementing each participant's baseline step count by a scaled proportion indexed by the participant's random treatment assignment and time from baseline. The four groups correspond to the following percentage increase for a participant averaging 1000 steps per day at baseline ("base percentage increase"): 20% (active control group), 100% (moderate increase group), 200% (large increase group), or 400% increase (very large increase group). To avoid excessively high target step counts, the target percentage increase is proportionally lower for participants with baseline step counts greater than 1000 (equation 1).

[1] Target percentage increase = base percentage increase * (baseline step count / 1000)^(-$\log_7 4$)

This target percentage increase is approached linearly from baseline to the end of week 6 (equation 2). Beyond week 6, the daily step count target remains stable.

[2] Daily step count target = baseline step count * (1 + target percentage increase * ([increasing week number between 1 (week 1) and 6 (week 6 and beyond) / 6]))

For participants averaging 1000 to 7000 steps per day at baseline, the target daily step count for the active control group is a 5-20% increase, for the moderate group a 25-100% increase, for the large group a 50-200% increase, and for the very large increase group a 100-400% increase (with smaller percentage increases for participants averaging more steps at baseline, Figure 3). The 5-20% increase is considered an active control group given that a step count increase of this magnitude is expected not to be clinically meaningful.

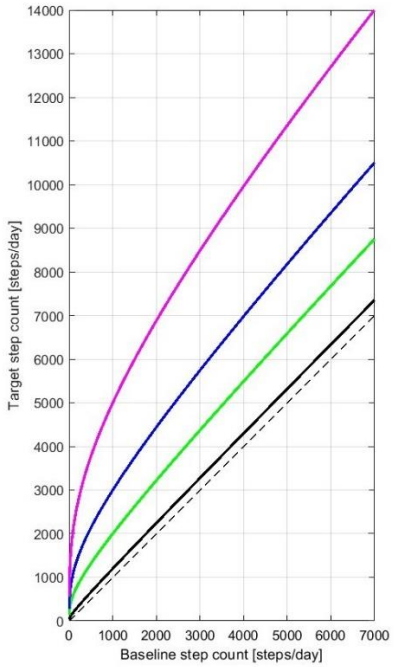

Figure 3. Target- and baseline step count. Dashed black line: line of identity. Black: active control group, green: moderate increase, blue: large increase, purple: very large increase.

2.4. Blinding

The study is double-blind meaning that the participants and the researchers are blinded to group allocation. The randomization is entered in the back-end of the app, whereafter participants have full access to the app. The app looks similar for participants in all groups to ensure blinding. Participants are unaware of the details of the allocation options: we tell participants that they will be randomized to one of four groups that are all motivated to take more steps, but to a different degree. Participants are able to view their absolute step counts per day (with only a history of one day) and their percentage of step count target reached for that week. The blinding is not broken prior to data lock unless the accredited medical research ethics committee (MREC) requests this.

2.5. Allocation concealment

Treatment assignment is performed by dr. Nienke de Vries (project leader), who is not involved in the intervention or data collection. In absence of dr. de Vries, a scientist, not involved in the intervention or data collection, performs treatment assignment. Group allocation is concealed for all other members of the study team.

2.6. Assessment of blinding success

Blinding of participants is checked at the follow-up assessment at one year by asking participants whether they think were randomized to a group with a small or a large increase in step count.

2.7. Stratification factors

Randomization is stratified by sex (two groups: female and male) and disease duration (three groups: <5 years, 5-10 years, and >10 years disease duration)

2.8. Sample size

The planned sample size is 452 participants. This sample size is based on a previous study of one-year change in step counts in a clinical trial evaluating the use of a smartphone application to increase physical activity of patients with chronic obstructive pulmonary disease (COPD) (Vorrink et al., 2016). Vorrink et al. (2016) reported a person-to-person standard deviation (SD) of change in one-year step count of 1957 in the active arm and 1973 in the control arm (Romeo et al., 2019). With 452 participants randomized 1:1:1:1 to the treatment arms (active control or moderate, large, or very large increase group), assuming an SD of 2000 steps and allowing for up to 20% loss to follow-up, the study will have greater than 90% power to infer a significant increase in step counts over one year if the expected 52-week increase in steps in the very large increase group relative to the active control group is at least 1000 steps based on a two-tailed test at $p < 0.05$ for this single primary comparison. One thousand steps is within the range of increases associated with exercise interventions among older adults and those with disabilities and chronic illness (Tudor-Locke et al., 2011).

2.9. Sample size review

No formal sample size re-estimation will be performed.

2.10. Framework

We will apply superiority hypothesis testing, comparing the effect of dose-dependent encouragement. We will compare the three interventional arms to the active control group.

2.11. Timing of final analysis

The final analysis will take place when the last follow-up visit (52 weeks post randomization) of the last participant, is performed and the data has been locked.

2.12. Schedule of assessments

See Table 3.

2.13. Timing of outcome assessments

The baseline visit is scheduled approximately 2 weeks after the baseline screening of step counts, if a person seems provisionally eligible. If the baseline visit cannot take place within 8 weeks after the step count screening, a new screening is performed.

The follow-up visit is scheduled 53 weeks after the baseline visit (52 weeks after randomization) and takes place no more than 8 weeks after the initial scheduled date. The mean and range of weeks from the baseline visit will be reported. Participants come in their regular medicated state.

Questionnaires are sent after the baseline and follow-up visit. Reminders are sent until one month after the initial questionnaire invitation.

Assessment	Pre-baseline screening	Baseline visit (T=0)	During intervention	Follow-up visit (T=12)	Continued until trial completion	52 weeks after follow-up visit
Participant characteristics						
Demographics		x				
Height (m)		x				
Weight (kg)		x		x		
Parkinson's disease characteristics						
Confirmed PD diagnosis, disease duration		x				
PD medication		x		x		
Onset symptoms		x				
If applicable: change in diagnosis confirmed by neurologist				x		
Physical activity level						
Step count with STEPWISE app (primary outcome)	x (4 wks)		x		x[#]	
Physical activity with Axivity AX6		x (1 wk)		x (1 wk)		
Self-reported physical activity level (LAPAQ)		x		x		x
Physical fitness						
6 minute walk test (6MWT)		x	x	x		

Cardiopulmonary exercise test (subgroup of 100 people)		x		x		
Remote cardiorespiratory fitness with mPower app **		x	x (1x/3mo)	x		
Motor symptoms						
Hoehn and Yahr stage		x		x		
Timed Up and Go Test (TUG)		x		x		
Mini-BestTest		x		x		
MDS-UPDRS III (in regular medicated state)		x		x		
MDS-UPDRS IV		x		x		
10MWT (gait speed)		x		x		
Handgrip strength		x		x		
Falls Efficacy Scale - International (FES-I) *		x		x		x
Lichamelijke Vaardigheden Schaal (LIVAS) *		x		x		x
Falls and near falls		x	x (monthly)	x		
Motor functioning with mPower app **		x	x (14 days/3mo)	x		
Non-motor symptoms						
MDS-UPDRS Ia, Ib* and II *		x		x		x
Montreal Cognitive Assessment (MoCA)		x		x		x

Hamilton Anxiety and Depression Scale (HADS) *		x		x		
Fatigue Severity Scale (FSS) *		x		x		x
Abbreviated version of the Apathy Evaluation Scale (AES-12PD) *		x		x		x
Scales for Outcomes in Parkinson's Disease (SCOPA) *		x		x		x
Quality of life						
Parkinson's Disease Questionnaire (PDQ-39) *		x		x		x
Blood-based biomarkers * **						
2 Serum samples (8.5 mL)		x		x		
2 Plasma samples (10 mL)		x		x		
1 Plasma sample for genotyping (6 mL)		x				
1 Whole-blood sample (6 mL)		x		x		
1 cell-free DNA sample (10 mL)		x		x		
Other						
Custom questionnaire on blinding				x		
Global Perceived Effect (GPE) *		x		x		
Self-reported treatment by physiotherapist		x		x		
Custom questionnaire on barriers and motivators to engage in physical activity		x		x		x

Custom questionnaire on use of the STEPWISE app		x		x		x
System usability Scale		x		x		x

Table 1. Overview outcome measures and timing. * These outcome measures will be collected with questionnaires which participants will fill in at home. ** Optional outcome measures, if participants do not want to perform this assessment they can still participate in the study. *** The blood withdrawal is optional for the last ~135 participants. # Only collected if participants opt-in to continue using the app after the 53-week follow-up.

3. Statistical Principles

3.1. Confidence intervals, p values and level of statistical significance

We will report 95% confidence intervals (CI), and nominal p-values. For the primary estimate, a $p<0.05$ is considered statistically significant. For all other estimates, we will report unadjusted p-values.

3.2. Adjustment for multiplicity

We use a hierarchical testing framework and refer to chapter 6 for details.

3.3. Adherence and protocol deviations

3.4. Definition adherence

We disentangle two types of adherence. Firstly, we will determine adherence to the intervention, which is the adherence of the participant to their step count target. Second, we will determine adherence to the use of the STEPWISE app, which is the frequency participants use the STEPWISE app.

3.4.1. Presentation adherence

For adherence to the step count target, we will summarize the mean steps per day during baseline and each week of the intervention, change in step count from baseline to every four-week period during follow-up, compliance to the step count target in every four-week period and relative increase in step count (presented in more detail in section 5.1).

For adherence to the STEPWISE app, we will summarize:

i) How often participants opened the app during the intervention (mean frequency/week of opening the app), summarized over all participants and per group

ii) How often participants opened the app *after* the intervention (mean frequency/per week of opening the app), summarized over all participants and per group. These data are only presented for those participants that provided additional consent for this part of the study.

3.5. Definition protocol deviations

Protocol deviations are any deviations from the protocol for a given subject. These are described in a Protocol deviation form. We distinguish the following types of deviations:

- Minor protocol deviation: when it is not expected that the event will have any significant impact on the quality of the data (primary outcome) or on the legal rights, safety, or privacy of the participant(s)
- Major protocol deviation: when the event may have an impact on the quality of the data (primary outcome) or on the legal rights, safety, or privacy of the participant(s)
- Protocol violation: events that definitely have an impact on the quality of the data (primary outcome), or on the legal rights, safety, or privacy of the participant(s). Protocol violations are reported to the ethic committee.

Other significant notes are stored in a Note to File form.

3.6. Presentation protocol deviations

A list of protocol violations will be presented with brief description.

4. Trial population

4.1. Screening data

Mean, standard deviation and range of step counts of participant excluded after the baseline period will be provided, as well as a list of reasons of exclusion (Table 2).

4.2. Eligibility

Inclusion criteria: having a confirmed PD diagnosis according to the MDS criteria (Postuma et al., 2015) by a neurologist, Hoehn and Yahr stage 1–3 during the clinical evaluation at baseline, being able to walk independently inside the home without the use of a walking aid, being able to understand the Dutch language, performing at most limited volume of physical activities prior to inclusion (i.e., taking fewer than 7,000 steps/day).

Exclusion criteria: if they have experienced weekly falls in the three months before enrollment, report medical conditions that hamper mobility other than PD, are not living independently, have cognitive impairments that hamper the use of a motivational app, or do not have a suitable smartphone (iPhone 5S or newer with iOS (iPhone Operating System) 10 or higher or Android 4.1 or newer).

The step count eligibility criterion is checked during a four-week baseline period. If interested participants meet this criterion, they are invited to visit Radboudumc to be further assessed for inclusion.

The CONSORT 2010 (Schulz et al., 2010) flow-chart will be completed. The information that will be reported is shown in Table 2.

Phase of trial	**Total**	**Active control group**	**Moderate group**	**Large group**	**Very large group**
Enrollment					
Assessed for eligibility	…				
Excluded	…				
Not meeting eligibility criteria	…				
No PD diagnosis	…				
HY >3	…				
Not able to walk independently inside	…				
Unable to understand Dutch	…				
≥7,000 steps/day at baseline	…				
Weekly falls past 3 months	…				
Conditions hampering mobility other than PD	…				
Not living independently	…				
Cognitive impairments hampering app use	…				
Not in possession suitable smartphone	…				
Declined to participate	…				
Other reasons	…				

Randomized	…	…	…	…	…
Allocation	…	…	…	…	…
Received randomization	…	…	…	…	…
Did not receive randomization	…	…	…	…	…
Follow-up					
Lost to follow-up	…	…	…	…	…
Reason 1	…	…	…	…	…
Reason 2	…	…	…	…	…
Reason …	…	…	…	…	…
Discontinued intervention	…	…	…	…	…
Reason 1	…	…	…	…	…
Reason 2	…	…	…	…	…
Reason …	…	…	…	…	…
Analysed	…	…	…	…	…
Excluded from analyses	…	…	…	…	…
Reason 1	…	…	…	…	…
Reason 2	…	…	…	…	…
Reason …	…	…	…	…	…

Table 2. CONSORT variables for progress through phases of clinical trial.

4.3. Recruitment

Participants are recruited in the Netherlands, although Dutch understanding participants from neighboring countries are also deemed eligible, provided that they are able to come to the Radboudumc. Participants are recruited using multiple strategies. First, we invite people with PD who are registered on the ParkinsonNEXT platform (n = 2,884; www.parkinsonnext.nl), which is an online platform that connects people with PD who are interested to participate in research with

researchers and clinical studies. Second, we advertise the study on social media (Facebook, Twitter, LinkedIn, Instagram) and on the website and newsletter of the Parkinson Vereniging (Dutch association for people with PD). We also recruit through our outpatient clinic (neurologists and PD nurse specialists), via referrals from specialized physiotherapists who are part of the national ParkinsonNet (network of allied health professionals working with PD) and by visiting Parkinson cafes (informative get-togethers for people with PD). Interested participants sign-up via www.parkinsonnext.nl/stepwise.

4.4. Withdrawal/Loss to follow-up

Withdrawal and loss to follow-up data will be presented in the CONSORT flow diagram. The data is collected upon notification by the participant. We distinguish between the following reasons for early trial discontinuation: loss to follow-up, withdrawal of consent, decision of the investigator, choice of the participant (in this case, the participant completes the follow-up visit), change in diagnosis, technical problems, (serious) adverse event, death or other.

5. Outcome definitions

5.1. Step count

The main outcome measure is change in step count per day as measured with the participant's smartphone. Step counts are collected on participants' smartphone with the phone's native algorithm. The STEPWISE app uses number of steps collected via a platform on iOS phones (HealthKit platform) and Android phones (Google Fit platform). These platforms collect the step count data on the background and are by default on all iPhones (5S or newer) and Android. The HealthKit platform is by default on every iPhone 5S or newer. The HealthKit platform collects data on physical activity in the background through the motion-chip. These data are locally (on the smartphone itself) saved by HealthKit and read by the STEPWISE app. The Google Fit platform is by default on most of the recent Android smartphones (version 4.1 and higher). The STEPWISE app reads the step counts from the Google Fit platform when the STEPWISE app is activated. Step counts are calculated per day from the accelerometer signal that is in all iPhone and Android smartphones. Different smartphones may well have different algorithms to calculate step counts, and we accept that there may be small differences in absolute step count numbers across devices. However, our aim here is to titrate the activities relative to a proportional (relative) increase in step

counts, regardless of how it was measured exactly by the device itself. This proportional (relative) increase in step counts is comparable across devices.

Step counts are collected per day and stored in the back-end of the app. Per participant, one .csv file with the columns 'ID', 'Date (Year-Month-Day)' and 'Steps per day' is exported after the participant finishes the study.

The following summaries of the step count data will be made, for the total sample and for each treatment group:

1. Mean steps per day during baseline period: mean steps per day over the four-week screening period. This step count is used as screening, but will also be used to analyze treatment effects.
2. Mean steps per day for each week during the intervention: averaging steps per day over seven days, starting directly after randomization. This yields 52 weekly step counts during the intervention.
3. Mean steps per day for each four-week period during the intervention: averaging steps per day over 28 days, starting directly after randomization. This yields 13 four-week step counts during the intervention (week 1-4, 5-8, 9-12, 13-16, 17-20, 21-24, 25-28, 29-32, 33-36, 37-40, 41-44, 45-48, 49-52).
4. Change in step count from baseline to every four-week period during follow-up.
5. Compliance to the step count target in every four-week period during follow-up: steps per four-week period divided by target step count, expressed as percentage.
6. The number of weeks each participant reached their target step count.
7. Relative increase in step count: steps per day in every four-week period during follow-up divided by baseline step counts, expressed as percentage.

Larger absolute step counts indicate more physical activity, larger step count changes indicate a larger change in physical activity.

If step counts are missing, the average steps/day for a given four-week period is calculated as the average over the days with step count data.

5.2. Six minute walk test (6MWT)

The six minute walk test (Butland et al., 1982) is a valid instrument to assess physical fitness. Participants walk over a course of 25 meters for six minutes. The number of times a participant

completes the course is multiplied by 25 and the number of meters the participant walked on the last course is added, yielding the total distance covered during the 6MWT.

Within-participant change in physical fitness is calculated by subtracting baseline distance covered from follow-up distance covered. Higher scores indicate better functioning. The 6MWT is performed at baseline and follow-up.

5.3. Parkinson's Disease Questionnaire (PDQ-39)

The PDQ-39 is the most widely used health related-QoL instrument in PD and is among the recommended scales to assess health related-QoL in PD (Martinez-Martin et al., 2011). The PDQ-39 (Peto et al., 1995) asks 39 questions organized over eight domains (scales): mobility (10 items), activities of daily living (6 items), emotional well-being (6 items), stigma (4 items), social support (3 items), cognition (4 items), communication (3 items), and bodily discomfort (3 items). Each item has five possible ordinal responses, from never to always, depending on frequency of the symptom over the preceding month. The eight scales' scores are generated by Likert's method of summated ratings and then transformed to a single total score that ranges from 0 to 100. Higher scores are associated with more symptoms. We will analyze the subdomains and the PDQ-39 total score. Participants complete the PDQ-39 at home after the baseline and follow-up visit through CastorEDC.

5.4. Movement Disorders Society Unified PD Rating Scale (MDS-UPDRS)

The MDS-UPDRS (Goetz et al., 2008) assesses PD symptoms and is administered at the baseline and follow-up visit.

The instrument is divided into four parts:

- Part I (non-motor experiences of daily living), comprising
    - Part IA concerning behaviors that are assessed by the Site Investigator with all pertinent information from participants and caregivers
    - Part IB that is completed by the participant with or without the aid of the caregiver, but independently of the Site Investigator.
- Part II (motor experiences of daily living), a self-administered questionnaire like Part IB
- Part III (motor examination) has instructions for the Site Investigator to give or demonstrate to the participant; it is completed by the Site Investigator.

- Part IV (motor complications) with instructions for the Site Investigator and also instructions to be read to the participant. This part integrates participant-derived information with the rater's clinical observations and judgments and is completed by the Site Investigator.

The full MDS-UPDRS has sixty-five items, each assessed on a 5-point Likert scale ranging from 0 to 4 with 0=none, 1=slight, 2=mild, 3=moderate, 4=severe. Total scores for Parts I, II, III, and IV and for Parts I through III collectively are calculated as simple sums of component items with mean imputation by Part if no more than 1, 2, 7, or 0 items is missing for Parts I through IV, respectively (Goetz et al., 2015). If more items are missing, the part is considered missing. Two additional summary scores will be constructed: patient-reported motor and non-motor aspects of experiences of daily living (sum of Parts IB and II) (Zou et al., 2023) and ambulatory capacity (sum of 5 MDS-UPDRS questions: walking and balance [question 2.12], freezing [q. 2.13], gait [q. 3.10], freezing of gait [q. 3.11], and postural stability [q. 3.12]). Higher scores imply worse symptoms.

Participants will self-administer Parts IB and II through CastorEDC. Parts IA, III and IV will be conducted by the Site Investigator during the in-clinic visits (Table 1). We strive to assess subjects by the same Site Investigator at baseline and follow-up.

5.5. Cardiopulmonary exercise test (CPET)

A subgroup of 100 participants perform a cardiopulmonary exercise test on a cycle ergometer at Canisius Wilhelmina Ziekenhuis. The CPET is performed after the regular visit to the Radboudumc, at baseline and follow-up.

During the CPET, the following parameters are collected:

- Oxygen consumption (mL/kg/min) in rest, at ventilatory threshold 1 and 2, at maximal capacity, after 1 minute recovery, after 3 minutes recovery
- Heart rate (beats/min) in rest, at ventilatory threshold 1 and 2, at maximal capacity, after 1 minute recovery, after 3 minutes recovery
- Lactate (mmol/L) in rest and at maximal capacity
- Power (Watt) at ventilatory threshold 1 and 2, at maximal capacity
- Respiratory Exchange Ratio (RER): in rest and at maximal capacity
- BORG (rate of perceived exertion, RPE 6-20) in rest and at maximal capacity. Higher score means higher perceived exertion

- Maximal exertion reached according to sports physician yes/no
- Reason for stopping the CPET: tired, pain in legs, shortness of breath, pain in chest, dizziness, other namely
- Observed chronotropic incompetence by sports physician yes/no
- Referral to cardiologist yes/no

Within-participant change from baseline in VO_{2max} in ml/kg/min is calculated by subtracting VO_{2max} at baseline from VO_{2max} at follow-up. Higher scores indicate better function.

5.6. Ten meter walk test (10MWT)

To determine comfortable and maximal gait speed, the ten meter walk test (10MWT) was performed (Collen et al., 1990). Participants walk over a course of 10 meters, with roughly 2 meters before and after the 10 meter walk course that do not count for determination of gait speed. Participants are given three attempts. The time (seconds) a participant takes for one attempt is entered to one decimal by the Site Investigator. The mean comfortable and maximal gait speed (m/s) is calculated by subtracting the distance (10 meters) by the time for completion (seconds). Moreover, the mean cadence (number of steps needed to complete the 10 meters) for comfortable and maximal speed is calculated (distance/number of steps).

Within-participant change from baseline in gait speed to follow-up is calculated by subtracting gait speed at baseline from gait speed at follow-up. Higher scores indicate better function.

5.7. MiniBESTest

Dynamic balance is assessed with the MiniBESTest (Franchignoni et al., 2010). The MiniBESTest consists of fourteen tasks that are scored on a three-point scale (0: severely impaired; 1: moderately impaired; 2: normal). The tasks are divided into four categories, for which the scores are summed:

- Anticipatory Postural Adjustments (sit to stand, rise to toes, stand on 1 leg (only worst side is scored)): range 0-6 points
- Postural Responses (stepping forward, backward and lateral (left and right, worst side is scored)): range 0-6 points
- Sensory Orientation (stance – eyes open; foam surface – eyes closed; incline – eyes closed): range 0-6 points
- Balance during Gait (gait during change speed, head turns, pivot turns, obstacles, cognitive "Get Up and Go" with dual task): range 0-10 points

The total score is calculated by adding the subscores of the four categories (range 0-28 points). If an item is missing, the subscore and total score is considered missing.

Within-participant change from baseline to follow-up is calculated by subtracting total MiniBESTest scores at baseline from MiniBESTest at follow-up. Higher scores indicate better function.

5.8. Hoehn and Yahr stage

The Hoehn and Yahr scale (Hoehn & Yahr, 1967) is used to stage PD motor manifestations and disability. The Hoehn and Yahr stage is scored by the Site Investigator at the end of the MDS-UPDRS-III evaluation (baseline and follow-up). Scores are on a ordinal scale and range from 0 to 5 with higher scores associated with more motor symptoms and disability. Stage 0 is "no signs of disease", stage 1 is "unilateral disease', stage 2 is "bilateral disease, without balance impairment", stage 3 is "mild to moderate bilateral disease; needs assistance to prevent falling on pull test", stage 4 is "severe disability, but still able to walk or stand unassisted" and stage 5 is "wheelchair bound or bedridden unless aided."

5.9. Handgrip strength

Handgrip strength is associated with mortality, cognitive decline, mobility, functional status in community-dwelling population (Rijk et al., 2016). Isometric handgrip strength is measured with a Jamar digital handgrip strength device (kilograms). Participants perform three attempts with either upper limb. In between trials, participants have 15-20 seconds of rest. After three trials of one limb, the other limb is tested. The highest attempt is analyzed for both limbs. Higher scores indicate better function.

5.10. Falls and near falls

History of falls are predictive of future falls and collecting data on falls is, even though they are prone to underreporting due to the retrospective nature of fall diaries (Keus SHJ, 2014). In this study, participants monthly receive a questionnaire on falls in the preceding month. If participants note they have fallen, the frequency (during the past month) and circumstances (i.e. date, direction of the fall, fall to the ground or not (in that scenario: near fall), position after the fall, injury).

5.11. Falls Efficacy Scale – International (FES-I)

Fear of falling is self-administered at baseline and follow-up with the Dutch version of the Falls Efficacy Scale International (FES-I) which has good psychometric properties in PD (Jonasson et

al., 2017; Yardley et al., 2005). The FES-I consists of sixteen items, rated on a four-point ordinal scale (1: not at all concerned; 2: somewhat concerned; 3: fairly concerned; 4: very concerned). The total score is calculated by summing the scores of the individual items. The maximal score is 64. The maximal number of missing items is 4 (25%), if a questionnaire misses more than four items, the questionnaire cannot be used. If a person misses 4 or fewer items, the total score is calculated as follows: FES-I score=(total score of items completed/#items completed)*16. The total score should be rounded up to the nearest whole number to give the score for an individual. Higher scores indicate more fear of falling. Two different staging schemes can be applied: 1) Scores ranging from 16-22 indicates limited fear of falling, scores ranging from 23-64 indicates severe fear of falling; 2) Scores ranging from 16-19 indicates low fear of falling, scores ranging from 20-27 indicates moderate fear of falling and scores ranging 28-64 indicates high fear of falling (Delbaere et al., 2010). Participants self-administer the FES-I through CastorEDC.

5.12. Hospital Anxiety and Depression Scale (HADS)

The Hospital Anxiety and Depression Scale (HADS) (Zigmond & Snaith, 1983) is self-administered at baseline and follow-up. The HADS assesses anxiety and depression, specifically over the past four weeks. The questionnaire is self-administered and consists of 14 items (7 for depression, 7 for anxiety) that are scored on a four-point Likert scale (0-3). For six items the scale is positive (0-3) and for eight items the scale is negative (3-0). The total score for anxiety is the sum of the items 1, 3, 5, 7, 9, 11 and 13. The total score for depression is the sum of items 2, 4, 6, 8, 10, 12 and 14. Higher scores indicate more anxious or depressive symptoms. There is little guidance in how to handle missing items. In cancer survivor's, it is recommended to use the mean of the participant for population inference ('subject mean') and the 'subscale half mean' (subject's subscale mean if at least half of the items were answered) for analysis at the individual level (e.g. screening). Any imputation performed better than leaving a participant out of the analysis (complete-case analysis) (Bell et al., 2016). Missing items will be replaced by the mean of the non-missing items of that assessment (baseline or follow-up) of that participant (mean imputation). Participants self-administer the HADS through CastorEDC.

5.13. Fatigue Severity Scale (FSS)

The Fatigue Severity Scale (FSS) (Krupp et al., 1989) is self-administered at baseline and follow-up. The FSS is validated for both screening and rating fatigue severity (Friedman et al., 2010). The FSS consists of 9 items on a seven-point Likert scale (1=completely disagree; 7=completely

agree). The total score is the mean of all items, yielding a score between 1 and 7, where higher score indicates more fatigue. Participants self-administer the HADS through CastorEDC.

5.14. Montreal Cognitive Assessment (MoCA)

The MoCA (Nasreddine et al., 2005) consists of 8 clinician-administered cognitive tasks designed to screen for mild cognitive impairment. The MoCA assesses attention and concentration, executive functions, memory, language, visuoconstructional skills, conceptual thinking, calculations, and orientation. The MoCA was developed to be more sensitive than the MMSE to patients presenting with mild cognitive complaint and may be less prone to a ceiling effect (Zadikoff et al., 2008). One point is awarded for correct completion of each item of the visuospatial/executive function task (5 items), naming task (3 items), digit vigilance and tapping items of the attention task (3 items), the sentence repetition items of the language task (2 items), abstraction task (2 items), delayed recall task (5 items), and orientation task (6 items). One point is awarded for naming 11 or more words during the fluency item of the language task. Zero (none correct) to 3 (4 or more correct) points are awarded based on the number of correct subtractions by 7 starting at 100 in the attention task. One point is awarded if the participant has 12 or fewer years of education unless the score is already 30. Scores for each task are summed for a total score (range 0 to 30) with higher scores indicating greater cognitive capacity. If an item is missing, the MoCA sum score for that participant is considered missing.

5.15. Abbreviated version of the Apathy Evaluation Scale (AES-12PD)

The Apathy Evaluation Scale (AES-12PD) is a shorter version of the Apathy Evaluation Scale (AES) and is reliable tool for the assessment of apathy in people with PD (Stankevich et al., 2018). The questionnaire addresses apathy in the past 4 weeks. The AES-12PD consists of 12 statements that are rated on a four-point scale (1: not at all true; 2: slightly true; 3: somewhat true; 4: very true). Lower scores reflect more apathy. A cut-off of 25 is recommended as indicator of apathy. Participants self-administer the AES-12PD through CastorEDC at baseline and follow-up. If an item is missing, the AES-12 PD score for that participant is considered missing.

5.16. Scales for Outcomes in Parkinson's Disease (SCOPA)

5.16.1. Autonomic dysfunction (SCOPA-aut)

The Scales for Outcomes in Parkinson's Disease for autonomic symptoms (SCOPA-aut) is proven reliable and valid (Visser et al., 2004). The SCOPA-aut consists of 26 items assessing the following regions: gastrointestinal (7), urinary (6), cardiovascular (3), thermoregulatory (4),

pupillomotor (1), and sexual (2 items for men and 2 items for women) dysfunction. Participants respond on a ordinal four-point scale of never (0) to often (3). The sexual items and the item on use of a catheter have different response options. The total score is obtained by summing the answers up (except for the question on medication, which is open ended): question 1-23 for men, question 1-21, 24 , 25 for women. The total score ranges from 0 to 69, higher scores reflecting worse autonomic functioning. Participants self-administer the SCOPA-aut through CastorEDC at baseline and follow-up. If a participant misses more than 25% of their items on the SCOPA-aut, they will be excluded from analysis. If a participant misses less than 25% of their items, the missing items will be replaced by the mean of the non-missing items of that assessment (baseline or follow-up) of that participant (mean imputation). Missing values on items addressing sexual problems will be imputed by the median value of participants from the same gender and disease duration and age onset group. If only one of the two items on sexual problems is missing, the missing item will be replaced by the non-missing item (Visser et al., 2008).

5.16.2. Sleep (SCOPA-sleep)

The Scales for Outcomes in Parkinson's Disease for daytime sleepiness and nighttime sleep (SCOPA-sleep) is reliable and valid (Marinus et al., 2003). SCOPA-sleep consists of 6 items on daytime sleepiness and 5 items on nighttime sleep. The nighttime subscale addresses nighttime sleep in the previous month. The 5 items are rated on a ordinal four-point scale (0: nota at all bothered by a sleep problem; 3: a lot bothered by a sleep problem). The scores of the nighttime items are summed, yielding a maximum score of 15 points. The additional question on overall sleep quality (7-point scale) is not included in the subscore. The daytime subscale addresses daytime sleepiness in the previous month. The 6 items are rated on a ordinal four-point scale (0: never; 3: often). The scores of the daytime sleepiness items are summed, yielding a maximum score of 18 points. Higher scores reflect worse sleep. Participants self-administer the SCOPA-sleep through CastorEDC at baseline and follow-up. If a participant misses more than 20% of their items on the SCOPA-aut, their questionnaire will be excluded from analysis (Marinus et al., 2003). If a participant misses less than 20% of their items, the missing items will be replaced by the mean of the non-missing items of that assessment (baseline or follow-up) of that participant (mean imputation).

5.17.Lichamelijke Vaardigheden Schaal (LIVAS)

The Lichamelijke Vaardigheden Schaal (LIVAS) is the Dutch translation of the Perceived Physical Ability questionnaire that addresses how someone perceives their physical abilities (Bosscher et al., 1995). The LIVAS consists of 10 items rated on a five-point scale. The sum scores range between 10 and 50. Six items (1, 2, 4, 6, 7, 9) are recoded so that higher scores represent more positive physical self-efficacy beliefs. Participants self-administer the LIVAS through CastorEDC at baseline and follow-up. If one item is missing, it is replaced by the mean of the remaining items of that participant of that assessment. If two or more items are missing, the LIVAS score for that participant is considered missing (Bosscher et al., 1995).

5.18.Self-reported physical activity (LAPAQ)

The Longitudinal Ageing Study Amsterdam (LASA) Physical Activity Questionnaire (LAPAQ) is a valid and reliable to classify older adults' physical activity (Stel et al., 2004). The LAPAQ is used for screening participants (questions 5-28 only) and for evaluation of physical activity at baseline and follow-up. During screening, the questions are asked by a Site Investigator on the phone, the baseline and follow-up questionnaire are self-administered through CastorEDC. The LAPAQ covers the frequency and duration of physical activities a person performed in the preceding two weeks: walking outside, biking, gardening, light- and heavy household activities, and a maximum of two sports activities. The LAPAQ consists of 37 items in total. Participants fill in the duration to the closest minute. Missing data for frequency and duration of an activity is imputed by assigning the mean value for participants that reported that type of physical activity, separately for men and women. We will calculate the following scores:

- Time spent in walking-related activities (walking and walking tours) in minutes per week
- Time spent in aerobic activities (biking outdoor and indoor, swimming, running, rowing) in minutes per week
- Time spent in physical activities (sum of all activities) in minutes per week
- Average energy expenditure (kcal/day): sum of energy expenditure of all activities, calculated per activity. Metabolic Equivalent of Task (MET) scores take into account the intensity of different activities. One MET is equivalent to 1 kcal/kg/h. The MET scores will be applied according to the Older Adult Compendium of Physical Activities (Willis et al., 2024). For each activity, we will determine the METs as follows: METactivity * (frequency activity over 2 weeks / 14) * (duration activity in minutes / 60). Energy

expenditure is the sum of the MET scores of each activity. We don't include body weight, because we want to study the average energy expenditure independent of weight (Ainsworth et al., 2011).

5.19.Blood-based biomarkers

A subgroup of participants (maximum n=135) will be asked to provide blood samples at their visit to the Radboudumc (2 serum samples of 8.5 mL, 2 plasma samples of 10 mL, 1 whole-blood sample of 6mL and 1 cell-free DNA sample of 10 mL at baseline and follow-up). One tube of plasma (6 mL) will be drawn at baseline only for genotyping. In total, we will draw 59 mL at baseline and 53 mL at follow-up. Participants opt-in for the blood drawings. Assays will be described when known.

5.20.Global Perceived Effect (GPE)

The Global Perceived Effect is the opinion of the participant regarding their recovery. The GPE consists of one question: "All things considered, how satisfied are you with the results of your treatment (participation in the STEPWISE trial)? ". The question is answered on a 7-point scale (1: extremely satisfied; 7: extremely dissatisfied) (Hudak & Wright, 2000). Scoring a 1 or 2 is considered 'clinical improvement', while scoring a 3 (somewhat satisfied) is considered 'no change'. Participants self-administer the GPE through CastorEDC at follow-up.

5.21.System usability Scale

Usability of the app is assessed at follow-up using the self-administered Dutch version of the system usability scale (SUS) (Brooke, 1995). The SUS consists of 10 items rated on a five-point Likert scale (1: strongly disagree; 5: strongly agree). Each item's score contribution ranges from 0 to 4. For items 1,3,5,7,and 9 the score contribution is the rating by the participant minus 1. For items 2,4,6,8 and 10, the contribution is 5 minus the rating by the participant. To obtain the overall usability score, one multiplies the sum of the scores by 2.5. This yields a total SUS score that ranges between 0 and 100. Participants self-administer the GPE through CastorEDC at follow-up. If any of the items is missing, the SUS score for that participant is considered missing

5.22.Custom questionnaire on blinding

The presumed treatment group to which a participant was assigned is collected with a self-administered questionnaire after the follow-up visit. The proportion of participants guessing

correctly or incorrectly and the odds ratio of guessing the true randomization versus the wrong randomization will be evaluated for each class of respondent.

5.23. Custom questionnaire on barriers and motivators to engage in physical activity

Participants' barriers and motivators to engage in physical activity is assessed in a custom made questionnaire consisting of 5 parts through CastoEDC:

- History with exercise
    - Frequency of exercise *before* PD diagnosis (not at all, 1x/wk, 2x/wk, 3x/wk, 4-6x/wk, every day, multiple times per day, other namely)
    - Types of exercise *before* PD diagnosis (multiple options, list of sports)
    - Age at which participant started to exercise (never, from childhood (<10 yrs), as teenager (10-20 yrs), between 20-30 yrs, after 30s)
    - Did exercise pattern change *after* PD diagnosis (no change, less exercise, more)
- Motivation to exercise at the moment (scale 1-10)
- Statements about motivators to exercise (17 statements; yes, no, not applicable)
- Statements about barriers to exercise (27 statements; yes, no, not applicable)
- Most important motivators and barriers (list 3 most important motivators and barriers).

5.24. Custom questionnaire on the use of other fitness apps

At baseline and follow-up, we ask participants to report (in a self-administered questionnaire through CastorEDC):

- Whether they use apps to monitor physical activity (yes, no)
- Which fitness apps they use (Health app on phone, Google Fit, Apple Health, Ommetje, Strava, Runkeeper, Fitbit, Garmin, other namely)
- Whether they use the STEPWISE app to monitor physical activity (yes, no; follow-up only).

At follow-up, we ask participants during the in-clinic visit whether they connected the STEPWISE app to their smartwatch. Data will be described as frequencies (percentages).

5.25. Custom questionnaire on use of the STEPWISE app

We pose a range of questions on the satisfaction and preferences of participants using the STEPWISE app through CastorEDC. These data will be used for potential further development of the app. Data will be described as frequencies (percentages).

- Was the STEPWISE app of added value (yes, no)
- How was your physical activity pattern this year, compared to the year before participation in the study (a lot less, less, similar, more, a lot more)
- To what extent did the STEPWISE app contribute to this (not, little, much, very much)
- How often did you manage to use your step count target (never, sometimes, almost always, always)
- Which part of the STEPWISE app did you like most (step count target, virtual chat, no preference)
- What would you change to the app so that it might motivate you more (connect to smartwatch, contact with other participants, notify illness of holiday, collect data other than walking, scoreboard with other participants, other namely)
- Would you like to keep using the app after the study (yes, no)

We also ask participants to describe the STEPWISE app in three words. We will describe the themes that arise from these words.

5.26. Participant characteristics

The following characteristics are collected at baseline: age, sex, ethnicity, smoking status (yes, no), years of education, highest education, employment, living situation, height, weight, treatment by physiotherapist (yes, no; frequency and duration). At follow-up, weight and any change in health status, employment and living situation is collected.

5.27. Background questions on PD

We collect the following data on the background of the PD diagnosis at baseline.

- Diagnosis confirmed by a neurologist (yes, no)
- Disease duration (years)
    - Stratum disease duration (<5, 5-10, >10 years)
- Years since first symptoms (numeric)
- Most affected side (left, right, symmetric)

- Medication use for PD (yes, no)
    - List of medication (see 5.29)
    - Use of advanced medication (DBS, apomorphine, duodopa, thalamotomy, yes other, no)

At follow-up, any change in diagnosis and medication is recorded.

5.28. Levodopa Equivalent Daily Dosage

The levodopa equivalent daily dosage (LEDD) will be calculated using data from the concomitant medications log. The conversion from dopaminergic drugs other than carbidopa-levodopa will follow the recommendations by Jost et al. (2023). The LEDD is obtained by multiplying the total dose of levodopa (number of tablets per day * dose per tablet) with the conversion factors below:

1) Levodopa/carbidopa: 1.
2) Levodopa/carbidopa, extended release: 0.75
3) Liquid intestinal levodopa/carbidopa (duopump): 1.11
4) Levodopa/carbidopa/entacapone: 1.33
5) Levodopa/carbidopa/tolcapone: 1.5
6) Levodopa/benserazide: 0.85
7) Levodopa/benserazide, extended release: 0.5
8) Pergolide: 100
9) Pramipexol: 100
10) Ropinirole: 20
11) Apomorfine: 10
12) Rotigotine: 30.3
13) Bromocriptine: 10
14) Piribedil: 1

15) Entacapone: the LED of dopa-containing medications is multiplied by 0.33 (so multiply the levodopa dose obtained in 1 through 7).
16) Tolcapone: the LED of dopa-containing medications is multiplied by 0.5 (so multiply the levodopa dose obtained in 1 through 7).
17) Biperideen: 0 (anti-cholinergic)
18) Trihexyphenidyl: 0 (anti-cholinergic)
19) Rasagiline: 100
20) Selegiline, tablet: 10
21) Safinamide: add 150 mg to the LEDD, independent of dose.
22) Amantadine: 1
23) Mucuna pruriens: 0 (unknown)
24) Carbidopa: 0 (unknown)
25) Propanolol: 0, this is a beta blocker
26) Rivastigmin plaster (mg/day): 0
27) Prolopa: 1
28) Liquid intestinal levodopa/carbidopa/entacapon (pump): daily dose * 1.11 (morning dose) + daily dose * 1.46 (maintenance and extra doses).

6. Analysis methods

6.1. General considerations

6.1.1. Statistical software

All statistical analyses will be performed using SAS (SAS Institute, NC, USA) or R (R Foundation for Statistical Computing, Vienna, Austria).

6.1.2. Summary statistics

Data will be summarized with respect to disposition, demographics (age, sex, marital status, education, ethnicity), pre-treatment characteristics, physical activity (step counts), secondary outcomes and safety outcomes. Summary statistics for continuous variables will include the

number of subjects, the mean, median, standard deviation, and range. For categorical data, summaries will include counts and percentages.

6.1.3. Precision

Results will generally be reported to 3 significant figures. Percentages will generally be reported to 0.1 percentage points. P-values will be reported to two digits when greater than or equal to 0.095, to three digits when greater than or equal to 0.001 and less than 0.095, and as <0.001 for all smaller values.

6.1.4. Administration

A test set of tables and figures specified in the SAP will be produced prior to breaking the blind using a dummy randomization schedule. The SAP will be finalized and must be approved by the all parties listed on this SAP prior to the final lock of the trial data and breaking of the blind.

6.2. Analysis populations

6.2.1. Intention-to-treat (ITT)

The intention-to-treat (ITT) sample consists of participants who are randomized, classified according to their randomized treatment assignment. Participants determined to have been ineligible prior to randomization but only discovered after randomization, participants who never initiated the intervention, and observations made after premature permanent discontinuation of the intervention are included in this sample. The analyses for our primary- and secondary objective will be performed with the ITT sample.

6.2.2. As-prescribed

We will determine whether there is any mismatch between randomization (in CastorEDC) and actual treatment assignment (performed in back-end of the app). If there is no mismatch at all, the as-prescribed analysis is dropped. If there is a mismatch, the as-prescribed sample consists of participants according to their actual assignment. The as-prescribed sample will be used for supportive analyses of the primary- and secondary analyses of the primary aim, for the primary analysis of intermediary effects of the secondary aim and for the primary analysis of efficacy of the secondary aim. The results obtained with the As-prescribed sample will be compared to the results from the analysis with the ITT sample.

6.2.3. As-treated (AT)

The as-treated (AT) sample consists of participants who are eligible, randomized, and participated for at least 14 weeks (the time to reach a stable step count target plus four weeks). If a participant permanently discontinues the intervention, observations made after discontinuation will be excluded. The AT sample will be used as supportive analyses for the primary- and secondary analyses of the primary aim, for the primary analysis of intermediary effects of the secondary aim and for the primary analysis of efficacy of the secondary aim.

The AT sample will be used as supportive analyses and will inform us whether, in participants using the STEPWISE app, dose-dependent encouragement i) increases physical activity levels (primary aim), ii) yields an effect on 53-week change in secondary outcome measures (secondary aim) and iii) whether there is a dose-response relationship between absolute 52-week step count change and secondary outcomes (tertiary aim). The results obtained with the AT sample will be compared to the results from the analysis with the ITT, and if applicable as-prescribed, sample.

6.2.4. Subgroup analyses

The following subgroups will be considered:

- Sex (male, female)
- Age (both categorized as <65 years vs. ≥65 years and continuous)
- Hoehn and Yahr stage (categorized as I, II and II)
- BMI (categorized as underweight <18.5, normal 18.5-25, pre-obese 25-30, obese ≥30 kg/m^2 and continuous)
- Time since diagnosis (categorized as <5 years, 5-10 years and >10 years and continuous)
- Baseline physical activity (both using quartiles and continuous)

For each subgroup, the potential for differential benefit from physical activity will be tested by including subgroup, subgroup x time, and subgroup x time x group (randomization) interaction terms into the primary random-slopes model. A significant subgroup x time x group 3-way interaction in combination with significantly slower progression among members of a subgroup randomized to any of the three intervention arm vs. members of the same subgroup randomized to the active control group will be taken as evidence of differential benefit.

6.3. Baseline participant characteristics

The baseline descriptive characteristics will be reported overall and per treatment group for the following characteristics: age, sex, ethnicity, smoking status, years of education, employment, living situation, body mass index, disease duration, time since first symptoms, disease severity (Hoehn and Yahr, MDS-UPDRS scores), LEDD, and MoCA. Baseline characteristics will be summarized as counts and percentages or as means, medians, standard deviations, and ranges. Baseline participant characteristics will not be tested.

6.4. Interim analyses and criteria

The aim of the interim analysis was to investigate whether participants increase their step count per month. We analyzed the volume of steps per month and pulled the intervention arms together in this analysis. We did not plan to terminate the intervention based on this analysis, but investigated whether we needed to change the intervention/the app. The interim analysis was planned to be performed at the point at which the 100th participant had completed 3 months of follow-up, corresponding to a total volume of follow-up of 600 person-months. Cut-offs for revision of the intervention / app were:

1) If less than 40% of participants in the intervention arms increase their step count by more than 20%, we need to revise the app

2) if more than 40% of participants in the control arm increases their step count by more than 35%, we need to revise the app

3) if the researchers see any reason they will revise the app

6.5. Primary aim

6.5.1. Primary analysis of primary aim

The primary analysis of the primary endpoint will estimate the feasibility of the intervention to induce a sustained within-participant change in daily step counts over 52 weeks. We will use the intention-to-treat (ITT) sample to evaluate the between-group change in step counts, comparing the active control group and each interventional group (moderate, large, and very large increase). The mean daily step count in the four-week baseline period will be compared to the mean daily step count in the four weeks prior to the week 53 visit (weeks 49-52 of the intervention). We will analyze participants' mean daily step count during each 4-week interval starting with the 4 weeks prior to baseline and ending with the 4 weeks prior to the week 53 visit in a shared-baseline, mixed

model repeated-measures (MMRM) analysis. The model will include fixed terms of observation interval (14 terms), treatment group x post-baseline interval interaction (3 x 13 = 39 terms), age x pre/post-treatment interaction (2 terms), and disease duration x pre/post-treatment interaction (2 terms). Covariance among the within-participant repeated measures will be assumed to be unstructured. The primary estimate will be the one degree of freedom linear contrast tested at two-tailed $p < 0.05$ comparing change from baseline to the final four weeks of the intervention (week 49-52) between the group randomized to a very large increment in steps vs. the active control group.

Secondary comparisons of the primary analysis include comparing the moderate and large increase groups to (a) the active control group to determine whether smaller increments also result in measurable increases in physical activity and (b) the very large increase group to determine whether a ceiling effect is reached.

Residuals will be checked for normality and homoscedasticity. Similar assessments and influence statistics will be evaluated for the other models reported for this trial.

6.5.2. Additional analyses of primary aim

The first additional analysis will consider group-dependent changes in step counts from baseline to 4, 8, 12, 16, 20, 24, 28, 32, 36, 40, 44, 48, and 52 weeks of follow-up to identify temporal patterns of response to determine whether larger increments occur early that are not sustained to the final four weeks of the intervention.

In the second additional analysis, we will apply an alternative model that consists of cubic splines with knots every 6 weeks (at 6, 12, 18, 24, 30, 36, 42 and 48 weeks).

6.5.3. Alternative covariance structures

If the model with unstructured covariance fails to converge, we will apply a piecewise linear covariance structure with knots every 6 weeks (so at 6, 12, 18, 24, 30, 36, 42 and 48 weeks). If this model also fails to converge, we will simplify the piecewise linear covariance by only one knot at 6 weeks, after which the step count target remained stable. If his model also fails to converge, we will apply a compound symmetric covariance structure.

6.5.4. Correlation step count and app engagement

To study the engagement with the app and change in step counts, we will compute the correlation between a) the number of days a participant opened the app during the intervention, b) how often

a participant opened the app during the intervention (sum), c) how often a participant opened the chat, d) how often a participant interacted with any of the progress screens, e) how often a participant interacted the progress for today, f) how often a participant interacted the progress screen to review their progress towards their target and g) how often a participant interacted the progress screen for the sum of their steps during the intervention, and the change in step counts from baseline to 52 weeks.

6.5.5. Other supportive analyses

Other supportive analyses include investigating the:

- Distribution of step count changes within weeks (increase in step counts on week days vs weekend days).
- Effect of the following potential moderators: baseline step count, Hoehn and Yahr stage, sex, employment status at baseline, living situation at baseline, minutes spent in total self-reported physical- and sports activities at baseline, aerobic activity (biking on an ergometer, running, rowing, swimming) at baseline, fear of falling, and balance (MiniBesTest). We will visualize the effects of the moderators in a Forest plot. This analysis will allow us to determine whether the effect of the intervention differs between strata (eg. Male versus female, HY 1 versus HY 3 etc.). The model will include the additional terms “moderator”, “moderator x time” and “moderator x time x group”.

6.5.6. Multiplicity adjustments for primary aim

The interim analysis was descriptive and did not include statistical testing. The final primary analysis will therefore be tested at a two-sided $p < .05$. For the secondary analyses, we will report nominal p-values.

6.6. Secondary aims

Our secondary aim is to evaluate whether dose-dependent encouragement through the STEPWISE app yields an effect on secondary endpoints. We will use equivalent ITT analyses as we used for the primary analysis but with fewer observation intervals to estimate treatment-associated differences in one-year change in these secondary outcomes. Variables that are strongly right-skewed will be log-transformed prior to analysis, and estimates will be back-transformed for reporting.

We will adjust the secondary outcome measures for age, sex, disease duration and baseline step count. MDS-UPDRS-III will be adjusted for age, sex, disease duration, baseline step count and LEDD.

6.6.1. Primary analysis of *intermediary* effects secondary aim

The primary intermediary estimate will be the comparison of change over 53 weeks (52 weeks post randomization) in six minute walking distance between the interventional arms combined (moderate, large and very large increase) and the active control group.

The minimal clinically important difference (MCID) in 6MWD is unknown for Parkinson's disease, but is reported to lie between 14 and 30.5 meters in patients with other pathologies (chronic obstructive pulmonary disease, lung cancer etc.) (Bohannon & Crouch, 2017). The minimum detectable change is 82 meters in people with parkinsonism (Steffen & Seney, 2008).

6.6.2. Primary analysis of *efficacy* secondary aim

The primary efficacy estimate will be the comparison of change over 53 weeks in motor- and non-motor aspects of experiences of daily living (sum of MDS-UPDRS parts IB and II) between the interventional arms combined (moderate, large and very large increase) and active control group.

6.6.3. Secondary analyses of secondary aim

In the secondary analyses of the secondary aim, we apply the same analyses as described for the primary intermediary and efficacy analysis of our secondary aim (§6.6.1 and §6.6.2), with all secondary outcome measures listed in §1.3.

6.6.4. Other supportive analyses of secondary aim

Other supportive analysis includes investigating step count changes over different time courses (e.g. baseline to 18 weeks follow-up, baseline to 30 weeks follow-up), and the effect on secondary outcomes.

6.6.5. Multiplicity adjustments for secondary aim

For the secondary analyses, we will report nominal p-values.

6.7. Tertiary aim

To test whether there is a dose-response relationship between amount of physical activity and physical fitness and motor- and non-motor functioning, we will regress 53-week change in clinical outcomes against 52-week cumulative step count (area under the curve, AUC), adjusting for age, sex, disease duration, and baseline step count. We will evaluate the clinical relevance of the

association of the step count with clinical outcomes through estimated effect sizes based on the regression terms, including 95% confidence intervals. We will analyze the full sample of participants irrespective of group allocation.

6.7.1. Primary analysis of *intermediary* effects tertiary aim

For the primary analysis for intermediary effects of our tertiary aim, we will regress the 53-week change in six minute walking distance against 52-week cumulative step count.

6.7.2. Primary analysis of *efficacy* tertiary aim

For the primary efficacy analysis of our tertiary aim, we will regress the 53-week change in motor- and non-motor aspects of experiences of daily living (sum of MDS-UPDRS parts IB and II) against 52-week cumulative step count.

6.7.3. Secondary analyses of tertiary aim

Secondary analyses for our third aim include relating the change in step count at different time points (analyzed in 4-week blocks, so for example from baseline to the average step count during week 12-16) with change in secondary outcome measures. We will generate matrix scatter plots and compute the correlations between change in step count at different time points and change in secondary outcome measures.

6.7.4. Other supportive analyses of tertiary aim

Other supportive analyses include determining the threshold of physical activity leading to clinically relevant changes using a generalized additive model with each of the predictors above modeled as low degree of freedom monotonic splines. The 52-week increment in step count yielding a MCID will be interpolated from the spline for 52-week change in step count, with a confidence interval estimated by bootstrapping.

In another supportive analysis we will assess the same model as for the primary and secondary analyses of our tertiary aim, but in addition adjust for baseline VO_{2max} since this variable is only available in 100 participants.

6.7.5. Multiplicity adjustments for tertiary aim

For the tertiary analyses, we will report nominal p-values.

6.8. Primary safety aim

For our primary safety aim we will investigate whether participants assigned to any of three different intervention arms experience more falls, or other adverse events (AEs) or serious adverse events (SAEs), than the control group during the study period (53 weeks).

6.8.1. Definition Adverse Events (AE)

Adverse events (AE) are defined as any undesirable experience occurring to a participant during the study, whether or not related to the use of the motivational application. Adverse events for which the participant obtains a medical check-up and are reported spontaneously by the participant and adverse events observed by the Site Investigator during in-clinic visits will be recorded. We will ask participants to report falls using a monthly fall diary sent through CastorEDC (Castor, 2019). Falls are considered an adverse event.

6.8.2.Definition Serious Adverse Events (SAE)

A serious adverse event (SAE) is any untoward medical occurrence that results in death, is life threatening, results in hospitalization or prolongs an existing hospitalization, or results in persistent/significant disability or incapacity. Medical events that did not result in any of the outcomes listed above due to medical or surgical intervention but could have had these outcomes based on judgment of the investigator are SAEs. An elective hospital admission is not an SAE.

6.8.3. Presentation Serious Adverse Events

The incidence of (S)AEs will be summarized by the number of events of a given classification experienced by participants in each treatment group and by the proportion of participants experiencing such an event. Falls are reported every month and considered an AE. SAEs will be summarized in aggregate across all Common Terminology Criteria Adverse Events (CTCAE) terms.

Aggregate summaries of (S)AE grade will include characteristics of: (a) seriousness, (b) severity, (c) relatedness to the intervention, (d) action taken with intervention. For each level of a given SAE characteristic, summaries will include the number of events and proportion of participants for which that level of a characteristic was the worst they experienced.

6.9. Other analyses

6.9.1. Intercurrent Events (IE)

An intercurrent event (IE) is an event occurring after treatment initiation and can affect the interpretation or the existence of the measurements associated with the clinical question. Possible IEs considered in the STEPWISE trial are:

- Discontinuation of treatment (secondary, but not primary, outcome(s) collected at one-year follow-up)
- Adverse event resulting in (temporary) missing data
- Technical problems in the STEPWISE app resulting in (temporary) missing step count data
- Initiation of physiotherapy, exercise training, other types of therapy or intervention, during the intervention period
- Initiation of another intervention during the study period
- Adjustment of medication

6.10. Missing data

6.10.1. Handing of missing data

For the calculation of the descriptive statistics, participants with missing data will not be considered for the specific missing descriptive, unless otherwise specified. Missing data for rating scales/questionnaires whereby specific missing data instructions are available will be handled according to these instructions.

For the primary analysis, we will calculate the mean step count over every four-week (average over days with available data). As sensitivity analysis, we will apply a weighting function that includes a weighting factor of (1/the number of observations (days with available step count data)^2). Participants with missing baseline data for secondary outcome variables will be retained in the analysis.

6.10.2. Missing completely at random (MCAR)

When data is missing completely at random (MCAR), the missingness is unrelated of the dependent and independent variables.

6.10.3. Missing at random (MAR)

The planned mixed model yields estimates that are unbiased conditional on the observed scores under a missing at random (MAR) assumption.

6.11. Testing blinding

Descriptive statistics will be summarized for belief in assignment to a group with little extra physical activity or a lot extra (stratified by correct and incorrect responses). All intervention arms are grouped together as 'a lot extra'. The following predictors of treatment assignment are tested by logistic regression:

- Randomization
- Age at baseline
- Sex
- Baseline physical activity level (baseline step count)
- Time of enrollment (year)

7. References

Ainsworth, B. E., Haskell, W. L., Herrmann, S. D., Meckes, N., Bassett, D. R., Jr., Tudor-Locke, C., Greer, J. L., Vezina, J., Whitt-Glover, M. C., & Leon, A. S. (2011). 2011 Compendium of Physical Activities: a second update of codes and MET values. *Medicine and Science in Sports and Exercise*, *43*(8), 1575-1581. https://doi.org/10.1249/MSS.0b013e31821ece12

Bell, M. L., Fairclough, D. L., Fiero, M. H., & Butow, P. N. (2016). Handling missing items in the Hospital Anxiety and Depression Scale (HADS): a simulation study. *BMC Res Notes*, *9*(1), 479. https://doi.org/10.1186/s13104-016-2284-z

Bohannon, R. W., & Crouch, R. (2017). Minimal clinically important difference for change in 6-minute walk test distance of adults with pathology: a systematic review. *Journal of Evaluation in Clinical Practice*, *23*(2), 377-381. https://doi.org/10.1111/jep.12629

Bosscher, R. J., Van Der Aa, H., Van Dasler, M., Deeg, D. J. H., & Smit, J. H. (1995). Physical Performance and Physical Self-Efficacy in the Elderly:A Pilot Study. *Journal of Aging and Health*, *7*(4), 459-475. https://doi.org/10.1177/089826439500700401

Brooke, J. (1995). SUS: A quick and dirty usability scale. *Usability Eval. Ind.*, *189*.

Butland, R. J., Pang, J., Gross, E. R., Woodcock, A. A., & Geddes, D. M. (1982). Two-, six-, and 12-minute walking tests in respiratory disease. *Br Med J (Clin Res Ed)*, *284*(6329), 1607-1608. https://doi.org/10.1136/bmj.284.6329.1607

Castor, E. D. C. (2019). *Castor Electronic Data Capture*. Retrieved August 28, 2019 from https://castoredc.com

Collen, F. M., Wade, D. T., & Bradshaw, C. M. (1990). Mobility after stroke: Reliability of measures of impairment and disability. *International Disability Studies*, *12*(1), 6-9. https://doi.org/10.3109/03790799009166594

Delbaere, K., Close, J. C., Mikolaizak, A. S., Sachdev, P. S., Brodaty, H., & Lord, S. R. (2010). The Falls Efficacy Scale International (FES-I). A comprehensive longitudinal validation study. *Age Ageing*, *39*(2), 210-216. https://doi.org/10.1093/ageing/afp225

Ellis, T. D., Cavanaugh, J. T., DeAngelis, T., Hendron, K., Thomas, C. A., Saint-Hilaire, M., Pencina, K., & Latham, N. K. (2019). Comparative Effectiveness of mHealth-Supported Exercise Compared With Exercise Alone for People With Parkinson Disease: Randomized Controlled Pilot Study. *Physical therapy*, *99*(2), 203-216. https://doi.org/10.1093/ptj/pzy131

Franchignoni, F., Horak, F., Godi, M., Nardone, A., & Giordano, A. (2010). Using psychometric techniques to improve the Balance Evaluation Systems Test: the mini-BESTest. *J Rehabil Med*, *42*(4), 323-331. https://doi.org/10.2340/16501977-0537

Friedman, J. H., Alves, G., Hagell, P., Marinus, J., Marsh, L., Martinez-Martin, P., Goetz, C. G., Poewe, W., Rascol, O., Sampaio, C., Stebbins, G., & Schrag, A. (2010). Fatigue rating scales critique and recommendations by the Movement Disorders Society task force on rating scales for Parkinson's disease. *Mov Disord*, *25*(7), 805-822. https://doi.org/10.1002/mds.22989

Goetz, C. G., Luo, S., Wang, L., Tilley, B. C., LaPelle, N. R., & Stebbins, G. T. (2015). Handling missing values in the MDS-UPDRS. *Mov Disord*, *30*(12), 1632-1638. https://doi.org/10.1002/mds.26153

Goetz, C. G., Tilley, B. C., Shaftman, S. R., Stebbins, G. T., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stern, M. B., Dodel, R., Dubois, B., Holloway, R., Jankovic, J., Kulisevsky, J., Lang, A. E., Lees, A., Leurgans, S., LeWitt, P. A., Nyenhuis, D., . . . Movement Disorder Society, U. R. T. F. (2008). Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. *Mov Disord*, *23*(15), 2129-2170. https://doi.org/10.1002/mds.22340

Hoehn, M. M., & Yahr, M. D. (1967). Parkinsonism: onset, progression and mortality. *Neurology*, *17*(5), 427-442. https://doi.org/10.1212/wnl.17.5.427

Hudak, P. L., & Wright, J. G. (2000). The Characteristics of Patient Satisfaction Measures. *Spine*, *25*(24), 3167-3177. https://journals.lww.com/spinejournal/fulltext/2000/12150/the_characteristics_of_patient_satisfaction.12.aspx

Jonasson, S. B., Nilsson, M. H., & Lexell, J. (2017). Psychometric properties of the original and short versions of the Falls Efficacy Scale-International (FES-I) in people with Parkinson's disease. *Health Qual Life Outcomes*, *15*(1), 116. https://doi.org/10.1186/s12955-017-0689-6

Jost, S. T., Kaldenbach, M. A., Antonini, A., Martinez-Martin, P., Timmermann, L., Odin, P., Katzenschlager, R., Borgohain, R., Fasano, A., Stocchi, F., Hattori, N., Kukkle, P. L., Rodríguez-Violante, M., Falup-Pecurariu, C., Schade, S., Petry-Schmelzer, J. N., Metta, V., Weintraub, D., Deuschl, G., . . . Dafsari, H. S. (2023). Levodopa Dose Equivalency in Parkinson's Disease: Updated Systematic Review and Proposals. *Movement Disorders*, *38*(7), 1236-1252. https://doi.org/10.1002/mds.29410

Keus SHJ, M. M., Graziano M, Paltamaa J, Pelosin E, Domingos J, et al. (2014). European physiotherapy guideline for Parkinson's disease. *KNGF/ParkinsonNet, the Netherlands*.

Krupp, L. B., LaRocca, N. G., Muir-Nash, J., & Steinberg, A. D. (1989). The Fatigue Severity Scale: Application to Patients With Multiple Sclerosis and Systemic Lupus Erythematosus. *Archives of Neurology*, *46*(10), 1121-1123. https://doi.org/10.1001/archneur.1989.00520460115022

Marinus, J., Visser, M., van Hilten, J. J., Lammers, G. J., & Stiggelbout, A. M. (2003). Assessment of sleep and sleepiness in Parkinson disease. *Sleep*, *26*(8), 1049-1054. https://doi.org/10.1093/sleep/26.8.1049

Martinez-Martin, P., Jeukens-Visser, M., Lyons, K. E., Rodriguez-Blazquez, C., Selai, C., Siderowf, A., Welsh, M., Poewe, W., Rascol, O., Sampaio, C., Stebbins, G. T., Goetz, C. G., & Schrag, A. (2011). Health-related quality-of-life scales in Parkinson's disease: critique and recommendations. *Mov Disord*, *26*(13), 2371-2380. https://doi.org/10.1002/mds.23834

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., & Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J Am Geriatr Soc*, *53*(4), 695-699. https://doi.org/10.1111/j.1532-5415.2005.53221.x

Peto, V., Jenkinson, C., Fitzpatrick, R., & Greenhall, R. (1995). The development and validation of a short measure of functioning and well being for individuals with Parkinson's disease. *Qual Life Res*, *4*(3), 241-248. https://doi.org/10.1007/bf02260863

Postuma, R. B., Berg, D., Stern, M., Poewe, W., Olanow, C. W., Oertel, W., Obeso, J., Marek, K., Litvan, I., Lang, A. E., Halliday, G., Goetz, C. G., Gasser, T., Dubois, B., Chan, P., Bloem, B. R., Adler, C. H., & Deuschl, G. (2015). MDS clinical diagnostic criteria for Parkinson's disease. *Mov Disord*, *30*(12), 1591-1601. https://doi.org/10.1002/mds.26424

Rijk, J. M., Roos, P. R., Deckx, L., van den Akker, M., & Buntinx, F. (2016). Prognostic value of handgrip strength in people aged 60 years and older: A systematic review and meta-analysis. *Geriatr Gerontol Int*, *16*(1), 5-20. https://doi.org/10.1111/ggi.12508

Romeo, A., Edney, S., Plotnikoff, R., Curtis, R., Ryan, J., Sanders, I., Crozier, A., & Maher, C. (2019). Can Smartphone Apps Increase Physical Activity? Systematic Review and Meta-Analysis. *J Med Internet Res*, *21*(3), e12053. https://doi.org/10.2196/12053

Schootemeijer, S., de Vries, N. M., Macklin, E. A., Roes, K. C. B., Joosten, H., Omberg, L., Ascherio, A., Schwarzschild, M. A., & Bloem, B. R. (2023). The STEPWISE study: study protocol for a smartphone-based exercise solution for people with Parkinson's Disease (randomized controlled trial). *BMC Neurol*, *23*(1), 323. https://doi.org/10.1186/s12883-023-03355-8

Schulz, K. F., Altman, D. G., Moher, D., & the, C. G. (2010). CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMC Medicine*, *8*(1), 18. https://doi.org/10.1186/1741-7015-8-18

Stankevich, Y., Lueken, U., Balzer-Geldsetzer, M., Dodel, R., Graber-Sultan, S., Berg, D., Liepelt-Scarfone, I., Hilker-Roggendorf, R., Kalbe, E., Kaut, O., Mollenhauer, B., Reetz, K., Schaffer, E., Schmidt, N., Schulz, J. B., Spottke, A., Witt, K., Linse, K., Storch, A., & Riedel, O. (2018). Psychometric Properties of an Abbreviated Version of the Apathy

Evaluation Scale for Parkinson Disease (AES-12PD). *Am J Geriatr Psychiatry*, *26*(10), 1079-1090. https://doi.org/10.1016/j.jagp.2018.06.012

Steffen, T., & Seney, M. (2008). Test-retest reliability and minimal detectable change on balance and ambulation tests, the 36-item short-form health survey, and the unified Parkinson disease rating scale in people with parkinsonism. *Physical Therapy*, *88*(6), 733-746. https://doi.org/10.2522/ptj.20070214

Stel, V. S., Smit, J. H., Pluijm, S. M., Visser, M., Deeg, D. J., & Lips, P. (2004). Comparison of the LASA Physical Activity Questionnaire with a 7-day diary and pedometer. *J Clin Epidemiol*, *57*(3), 252-258. https://doi.org/10.1016/j.jclinepi.2003.07.008

Tudor-Locke, C., Craig, C. L., Aoyagi, Y., Bell, R. C., Croteau, K. A., De Bourdeaudhuij, I., Ewald, B., Gardner, A. W., Hatano, Y., Lutes, L. D., Matsudo, S. M., Ramirez-Marrero, F. A., Rogers, L. Q., Rowe, D. A., Schmidt, M. D., Tully, M. A., & Blair, S. N. (2011). How many steps/day are enough? For older adults and special populations. *Int J Behav Nutr Phys Act*, *8*, 80. https://doi.org/10.1186/1479-5868-8-80

Visser, M., Marinus, J., Stiggelbout, A. M., & Van Hilten, J. J. (2004). Assessment of autonomic dysfunction in Parkinson's disease: the SCOPA-AUT. *Mov Disord*, *19*(11), 1306-1312. https://doi.org/10.1002/mds.20153

Visser, M., van Rooden, S. M., Verbaan, D., Marinus, J., Stiggelbout, A. M., & van Hilten, J. J. (2008). A comprehensive model of health-related quality of life in Parkinson's disease. *Journal of Neurology*, *255*(10), 1580-1587. https://doi.org/10.1007/s00415-008-0994-4

Vorrink, S. N., Kort, H. S., Troosters, T., Zanen, P., & Lammers, J. J. (2016). Efficacy of an mHealth intervention to stimulate physical activity in COPD patients after pulmonary rehabilitation. *Eur Respir J*, *48*(4), 1019-1029. https://doi.org/10.1183/13993003.00083-2016

Willis, E. A., Herrmann, S. D., Hastert, M., Kracht, C. L., Barreira, T. V., Schuna, J. M., Jr., Cai, Z., Quan, M., Conger, S. A., Brown, W. J., & Ainsworth, B. E. (2024). Older Adult Compendium of Physical Activities: Energy costs of human activities in adults aged 60 and older. *J Sport Health Sci*, *13*(1), 13-17. https://doi.org/10.1016/j.jshs.2023.10.007

Yardley, L., Beyer, N., Hauer, K., Kempen, G., Piot-Ziegler, C., & Todd, C. (2005). Development and initial validation of the Falls Efficacy Scale-International (FES-I). *Age Ageing*, *34*(6), 614-619. https://doi.org/10.1093/ageing/afi196

Yerrakalva, D., Yerrakalva, D., Hajna, S., & Griffin, S. (2019). Effects of Mobile Health App Interventions on Sedentary Time, Physical Activity, and Fitness in Older Adults: Systematic Review and Meta-Analysis. *J Med Internet Res*, *21*(11), e14343. https://doi.org/10.2196/14343

Zadikoff, C., Fox, S. H., Tang-Wai, D. F., Thomsen, T., de Bie, R. M., Wadia, P., Miyasaki, J., Duff-Canning, S., Lang, A. E., & Marras, C. (2008). A comparison of the mini mental state exam to the Montreal cognitive assessment in identifying cognitive deficits in Parkinson's disease. *Mov Disord*, *23*(2), 297-299. https://doi.org/10.1002/mds.21837

Zigmond, A. S., & Snaith, R. P. (1983). The hospital anxiety and depression scale. *Acta Psychiatr Scand*, *67*(6), 361-370. https://doi.org/10.1111/j.1600-0447.1983.tb09716.x

Zou, H., Goetz, C. G., Stebbins, G. T., Schrag, A., Mestre, T. A., & Luo, S. (2023). Summing MDS-UPDRS Parts 1 + 2 (Non-motor and Motor Experience of Daily Living): The Patient's Voice. *Mov Disord*, *38*(7), 1363-1364. https://doi.org/10.1002/mds.29417