Statistical Analysis Plan of "STOP - Successful Treatment of Paranoia" Trial

STOP - Successful Treatment of Paranoia: Replacing harmful paranoid thoughts with better alternatives: A parallel randomised controlled trial of a digital therapy support app

Refers to: Study Protocol Version 3.2

SAP filename: STOP SAP_V1.2.docx

REC Reference: 21/LO/0896

Trial registration: ISRCTN17754650

Names, affiliations, and roles of SAP contributors:

Trial Statistician: Mr George Vamvakas, Department of Biostatistics and Health Informatics, IoPPN, King's College London

Senior Statistician: Prof Daniel Stahl, Department of Biostatistics and Health Informatics, IoPPN, King's College London

Chief Investigator: Prof Jenny Yiend, Department of Psychosis, IoPPN, King's College London

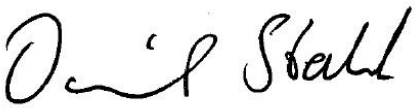
Project Manager: Tanya Ricci, Department of Psychosis, IoPPN, King's College London

Signatures

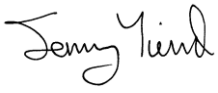
Trial Statistician: Mr George Vamvakas

Signature  Date: 07/11/2024

Senior Statistician: Prof Daniel Stahl

Signature…  Date: 07/11/2024

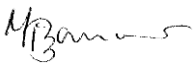
Chief Investigator: Prof Jenny Yiend

Signature  Date: 05/11/2024

Data Monitoring Committee Chair: Prof Andrew Gumley

Signature  Date: 05/11/2024

Trial steering Committee Statistician: Dr Mizanur Khondoker, on behalf of Trial Steering Committee Chair, Prof Jesus Perez

Signature  Date: 07/11/2024

CONTENTS

This document contains up to date statistical analysis plans (with version numbers and dates).

This SAP refers solely to the main clinical effectiveness analysis of the primary and secondary trial outcomes (i.e., the analysis to produce the main trial results paper and the secondary dose response analyses).

SAP REVISION HISTORY

Date	SAP version	Protocol version and date	Reason for change/amendment made
	v1.0	v1.01, 09/01/2022	Initial version
15/11/2022	v1.1	v3.2, 12/09/2022	Changes and amendments requested by MHRA
21/12/2022	V1.2	v3.2, 12/09/2022	Addition of CONSORT diagrams

Table of Contents

1 Trial background and rationale 5

2 Trial Design including blinding 7

3 Study setting 8

4 Study criteria 8

4.1 Eligibility criteria 8

4.2 Inclusion criteria 8

5 Randomization: Method of allocation of groups 10

5.1 Sequence generation 10

5.2 Concealment mechanism 11

5.3 Implementation 11

6 Sample size estimation 12

7 Interventions and frequency of assessments 12

7.1 Explanation for the choice of comparators 12

7.2 Intervention description 13

7.2.1 STOP Intervention 13

7.2.2 Text-reading control 13

7.2.3 Treatment as usual (TAU) 13

8 Measures 14

8.1 Baseline measures 14

8.2 Primary Outcome Measure 14

8.3 Secondary Outcome measures 15

8.4 Follow-up assessments 15

8.5 Interim assessments 15

9 Principal Research objectives 16

9.1 Estimand 16

9.2 Primary objective 16

9.3 Secondary Objectives 17

9.4 Exploratory Objectives 17

10 Intercurrent events and Treatment adherence, protocol violations, withdrawal from treatment and trial, and early trial stopping 18

10.1 Intercurrent Events 18

10.1.1 Adherence 19

10.2 Protocol violations 25

10.3 Withdrawal from treatment and trial 25

10.4 Early trial stopping 26

10.5 Loss to follow-up and other missing data 26
11 Adverse event reporting 26
11.1 Urgent reporting 26
11.2 Regular Reporting of serious adverse and adverse events 27
12 Data analysis plan 27
12.1 Baseline comparability of randomised groups 27
12.2 Descriptive statistics for outcome measures 27
12.3 Inferential analysis 28
12.4 Analysis of primary outcome 28
12.4.1 Model assumption checks 32
12.5 Analysis of secondary outcomes 32
12.6 Level of significance and methods for handling multiple comparisons 33
12.7 Missing data 34
12.7.1 Minimizing attrition at follow-up 34
12.8 Reporting of missingness 34
12.8.1 Missing data in baseline variables 35
12.8.2 Missing data in the outcome 35
12.8.3 Missing data in scales and subscales 39
12.9 Additional sensitivity analyses 40
12.9.1 Sensitivity of results to potential imbalance of baseline characteristics 40
12.9.2 Sensitivity of results to the increase in precision 40
12.9.3 Sensitivity analysis due to Protocol violations 41
12.10 Exploratory dose-response analyses 41
12.11 Planned subgroup analyses 44
12.12 Interim analysis 44
13 Software 44
14 Access to Protocol, participant level-data and statistical code 44
15 References 45

1 Trial background and rationale

Psychosis is one of the most disabling mental health conditions, with a lifetime rate of 3.5 (Perälä et al. (2007). It is associated with significant distress, increased unemployment, suicidal ideation, and impaired social functioning and physical ill-health (Freeman et al. (2011). Persecutory delusions—characterised by paranoid thinking—are the most frequent and clinically significant symptoms of psychosis. Researchers have shown that paranoid thinking in the general population is continuously distributed, indicating a hierarchical

structure to paranoia, ranging from social evaluative concerns (e.g., fears of rejection) to severe threat (e.g. being subject to significant harm) (Freeman et al. (2005); Bentall et al. (2009); Birchwood & Trower (2006); Garety et al. (2001); Savulich et al. (2015)). Persecutory delusions fall at the extreme point on the continuum of paranoid belief. As such, they are associated with more distress than other types of delusion (Freeman (2002)), are most likely to be acted upon (Wessely et al. (1993)) and represent a strong predictor of hospitalization (Castle et al. (1994)). Over one-third of all UK psychiatric patients suffer from persecutory delusions, often appearing in a range of psychopathologies, including depression (Johnson et al. (1991), bipolar disorder (Goodwin & Jamison (2007), posttraumatic stress disorder (Hamner et al. (1999), anxiety (Van O et al. (1999), and with the highest prevalence and greatest intensity in schizophrenia (Appelbaum et al. (1999).

The National Institute for Health and Care Excellence's (NICE) clinical guidelines recommend using Cognitive Behavioural Therapy (CBT) for treating psychosis. New directions in treatments for delusions emphasize briefer, targeted interventions, with a focus on putative causal factors such as cognitive biases (Moritz & Woodward (2007); Waller et al. (2011)).

Cognitive Bias Modification (CBM) techniques are a theory-driven treatment development that use a computerised task to manipulate biased interpretations and promote more adaptive processing. CBM-pa is a computer desktop version of this class of intervention specifically designed to target paranoid interpretations. Feasibility testing (Yiend et al. (2017) of a six-session CBM-pa has shown positive results (Yiend et al. (2017). Here, we aim to build on CBM-pa by testing a novel, entirely self-administered digital therapeutic for paranoia called STOP ('Successful Treatment of Paranoia'). STOP is the successor to CBM-pa and takes the form of a 12-session mobile app, adding six newly created sessions. Based on the feedback gleaned from the feasibility study of CBM-pa (Leung et al (2017), STOP includes illustrations with each item used in the study (i.e., graphics depicting the ambiguous scenarios and the non-paranoid interpretation that runs counter to the paranoid reader's initial assumption). Furthermore, STOP mitigates adherence issues commonly faced by online interventions by monitoring and rewarding participants with gamification techniques and scheduling protocols. Specifically, clinical researchers will collaboratively work with participants to

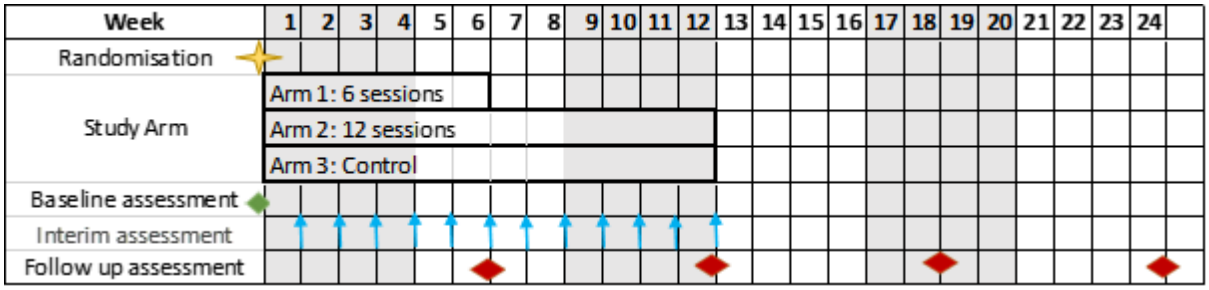
schedule interventions on the STOP mobile App; they will also have direct contact with participants during main assessment sessions or when participants miss a session.

2 Trial Design including blinding

We will conduct a four-year, parallel arm two-site superiority randomised controlled trial (RCT) with patients who experience clinical levels of paranoia. The trial is a three-arm RCT, where two doses of therapy, a 6- and a 12-session STOP mobile App therapy, are compared with a 12-session text reading control. We will randomly allocate (stratified by site and sex at birth) participants into one of the three trial arms. Hence, the treatment groups will be: 1) the 6-session intervention group (Arm 1), 2) the 12-session intervention group (Arm 2), and 3) the 12-session text reading control (Arm 3).

All three conditions will be conducted in addition to Treatment as Usual (TAU), described in section 7 “Interventions and Frequency of Assessments”.

Assessments will be taken at baseline and then at 6, 12 (end of treatment), 18-, and 24- weeks post-randomisation, with the primary outcome (severity of paranoid symptoms as measured by the Paranoia Scale, Fenigstein & Vanable (1992) set at 24 weeks – see Measures and Principal Research Objectives for details). Follow-up is anchored to randomisation. The use of a 6-week assessment permits an end of treatment assessment in Arm 1 and a mid-treatment assessment in Arm 2. The use of an 18-week assessment permits a 6-week and 12-week post-treatment comparison between all three arms. Data will also be collected at 12 weekly intervals during the treatment which permits additional, secondary, dose-response analyses alongside the traditional three-arm comparisons.



The trial statistician (GV) will report to the open and closed sessions of the DMC and will remain partially blind (he will see that some patients receive the same treatment but will not know which treatment it is) whenever possible until the main analyses are complete. The senior statistician (DS) will remain fully blinded until the analyses are complete. See Protocol for further details on blinding arrangements and procedures for unblinding (if needed). The trial will follow the Standard Operating Procedures of King's Health Partner's Clinical trials unit, see https://khpcto.co.uk/SOPs/00_SOPs.php

3 Study setting

This study will be conducted across two sites. The lead site is the Institute of Psychiatry, Psychology & Neuroscience (IoPPN), King's College London. The second site is the Department of Psychology, University of Bath.

4 Study criteria

4.1 Eligibility criteria

Eligible participants will be outpatients with distressing paranoid symptoms, irrespective of primary diagnosis. This is appropriate because STOP is designed to target distress caused by paranoid symptoms rather than a specific diagnosis. SCID-V Research version (SCID-5-RV) and PANSS clinical interviews will be conducted during baseline assessment to establish diagnoses and comorbidity. Potential participants will be screened and selected for invitation to participate by researchers according to the inclusion and exclusion criteria.

4.2 Inclusion criteria

1. Any clinically significant persecutory or paranoid symptoms, present for at least the preceding month. This will be operationalized as a score of 3 ("mild") or more on item 6 of the Positive and Negative Syndrome Scale (PANSS, [25]). See Protocol for details on PANSS and the scoring system.

2. Displaying an interpretation bias ≥ -2 on the 8-item screening version of the Similarity Ratings Task (SRT). See Protocol for details on SRT, the scoring system, and the choice of the cutoff.

3. If on psychotropic medication, then stable on that medication for at least the last 3 months and expected to be so for the study duration.

4. Age 18 years or over.

5. Capacity to consent; if in doubt assessed using the Capacity Assessment Tool.

For details and rationale, please see STOP Study protocol.

4.3 Exclusion criteria

1. Severe cognitive impairment.

2. Illiteracy or inability to understand written and spoken English for any other reason.

3. Major current physical illness (e.g., cancer, heart disease, stroke, dementia).

4. Major substance or alcohol misuse, assessed by the SCID-V screen.

5. Currently receiving, or soon due to receive, a psychological intervention targeting the same psychological mechanism as STOP (i.e., paranoid belief change) or having done so in the last 3 months.

6. Not currently taking part in any other interventional research study.

7. Scoring 7 (defined as 'Extreme - a network of systematised persecutory delusions denotes the patient's thinking, social relations and behaviour') on item 6 of the PANSS.

8. At high risk of suicide as indicated by the Columbia Suicide Severity Rating Scale (C-SSRS) - Screen Version [78,79]. This instrument uses a series of standardised questions covering ideation, intent, plans and behaviour over the last 1-3 months. Those falling into the high-risk category (i.e., expressing intent -with or without a plan- or reporting a suicidal behaviour) will be excluded and their referring clinician alerted.

Comment: Exclusion criteria 7 and 8 are introduced in order to limit the clinical risk associated with vulnerable patients taking part in a clinical trial of a 'Software as a Medical Device' (SaMD) product.

For details and rational, please see STOP Study protocol.

5 Randomization: Method of allocation of groups

5.1 Sequence generation

Randomisation is at the patient level and is performed using an online randomisation system set up King's Clinical Trials Unit (KCTU). The randomization procedure will be completed in the STOP app.

The KCTU system comprises a bespoke MACRO database which will be created in collaboration with the trial analysts and the CI and maintained for the duration of the project. It will be hosted on a dedicated server within KCL.

Researchers will be registered on the KCTU system and assigned usernames and passwords. They will log into the KCTU MACRO database system in order to generate a unique PIN for each participant at the point of referral to the trial (see implementation, below). Consecutive PINs will therefore encode both eligible and non-eligible participants, since consent is required prior to screening to collect screening related data. Participants randomised to the trial will therefore have non-consecutive PINs.

At the screening appointment, researchers will enter that PIN into Health Machine (HM; the researcher platform for administering the STOP app) which will register the participant onto the STOP trial. Baseline assessment will be completed during or after this first appointment using Health Machine and the STOP app.

Randomisation will take place at the end of baseline assessment using a checkbox within the HM platform to provide an exportable timestamp. At this point the PIN will be randomly allocated to one of the three study arms automatically by the HM/ STOP app software. KCTU will provide Avegen with randomised stratified lists ahead of the trial start date, to which PINs will be assigned consecutively at the time of randomisation.

Randomisation will be at the individual level in the ratio 1:1:1 to the three trial arms and performed independent of the study team and automatically within the STOP app, as described above.

Randomisation stratifiers (measured at screening) will be: 1) study site (Bath, London) and 2) sex at birth (male, female).

5.2 Concealment mechanism

In this study randomisation lists including participant PIN numbers will be generated and communicated by KCTU to Avegen. Avegen will be responsible for implementing the randomisation process within their software platform. The patient's assignment to study arm (1, 2 or 3) will remain concealed from all those involved in the study except two nominated individuals at Avegen who do not have any contact with the research team. These individuals will receive the KCTU randomisation lists and programme the HM/STOP app to consecutively assign PIN numbers to the randomisation list at the time the researcher completes the 'randomise' checkbox in HM.

All communications with patients about their assignment to condition (e.g., publicity, information sheet, in-app information etc.) will use the following wording: "You will be randomised to one of three different procedures, of varying lengths, some of which we think may help reduce symptoms of paranoia."

Neither patients nor researchers will be able to discern the difference between study arms since all will be procedurally identical and all will involve 12 scheduled sessions (the 6 arm intervention group will receive 12 in-app interim assessments, but without the preceding in-app therapy session)

5.3 Implementation

PIN allocation will be undertaken by trial researchers going to www.ctu.co.uk and clicking the link to access the MACRO system. A full audit trail of data entry will be automatically date and time stamped, alongside information about the user making the entry within the system. If there are any mistakes (e.g., randomising a screen failure in error) then it is important not to 'undo' it - errors should stay in the system. Randomisation errors will be communicated by Avegen to the statisticians via notes added by Avegen in their system. The CI or delegate (e.g., Trial Manager) may request Avegen to add notes against individual subject entries to clarify data entry errors related to randomisation.

See Figure 6 of the study protocol for a diagram on the randomisation procedure.

6 Sample size estimation

With alpha = .05 and 80% power, recruiting a sample of 77 per group would permit detection of a clinically useful drop of 8 points on the Paranoia Scale compared to TAU assuming a standard deviation of 17.5 (based on feasibility data). This magnitude drop would be a 15% reduction in the average score of paranoid patients (52: [8]) and represents an effect size of d=0.46 (small to medium). Assuming a conservative estimate of 15% drop-out (from feasibility data) a total sample size of 273 is required when using an independent t-test.

With a sample of 273 patients and 2 binary stratifiers we will have 136 and 66 in the first and second strata respectively, meaning we can randomize 66 people in 2 arms in each of the 4 strata combinations.

7 Interventions and frequency of assessments

7.1 Explanation for the choice of comparators

The study uses an active control condition as this directly evaluates the effect of specifically manipulating the hypothesised mechanism of action (bias reduction) and controls for any

non-specific therapy effects (e.g., contact with researchers, interaction with digital equipment, trial participation etc.) in a manner not possible with a waiting list control.

7.2 Intervention description

7.2.1 STOP Intervention

Participants randomised to this condition will receive 40 training items per session (approximately 40min duration) on the STOP mobile App. Participants will read text inviting paranoid interpretations, then complete missing words and answer questions in a way that encourages alternative, adaptive beliefs about themselves and others (see Protocol for an example). In addition to the existing 240 training items (40 training items across six sessions) from the CBM-pa feasibility study, we will create an additional set of 240 training items for the additional six sessions (i.e., sessions 7-12 for the 12-session arm). See Protocol for details.

7.2.2 Text-reading control

Participants randomised to the active control condition will also receive 40 control items per session (approximately 40min duration) on the STOP mobile App. The experience will be identical to the intervention condition except for the item content that participants see omits the active ingredient: resolution of an emotionally ambiguous situation in a benign/non-paranoid manner. Instead, control participants read and respond to factual material or mundane everyday experiences (see Protocol for an example).

7.2.3 Treatment as usual (TAU)

All three conditions will be conducted in addition to Treatment as Usual (TAU), which will be in the form of individualised combinations of medication and care coordination and may include eligible psychological therapies (i.e., those which do not target or interact with the same mechanism as STOP). We will record details of TAU at each main outcome assessment and changes on a weekly/ fortnightly basis during the study period using a standardised template to record participants' responses to a series of set questions about treatment received, including pharmacotherapy and any treatment changes.

See Protocol for details on any other concomitant care permitted or prohibited during the trial.

8 Measures

8.1 Baseline measures

The following measures are recorded at baseline: three relevant items from the Credibility/Expectancy questionnaire (CEQ): a six-item measure used to measure treatment credibility and expectancy in psychotherapy research (Devilly & Borkovec, 2000). IQ (WTAR). Persuadability subscale (PER). Paranoid beliefs (*SRT*, *SST*). Clinical symptoms assessed using PANSS Item 6 (P6 Suspiciousness/Persecution), Paranoia Scale, Green Paranoid Thought Scale Revised (R-GPTS), Paranoia Worries Questionnaire (PWQ), and Hospital Anxiety and Depression Scale (HADS), as well as the Structured Clinical Interview for DSM-5 (SCID-5) [47] (conducted by a trained team member during baseline assessment. Recovery measures). Recovery, measured using the Warwick-Edinburgh Mental Wellbeing Scale (WEMWBS), Questionnaire about the Process of Recovery (QPR), Recovering Quality of Life (ReQoL) questionnaire, and the Ben-Zeev et al. (2014) mobile app user satisfaction measure.

Socio-demographic characteristics recorded at baseline are: age, age of onset of distressing paranoia, gender, ethnicity, education level, employment, living arrangement, relationship status, self-reported dyslexia.

8.2 Primary Outcome Measure

The primary outcome is the severity of paranoid symptoms as measured by the Paranoia Scale (Fenigstein & Vanables). See protocol for details on the scale's items and scoring method.

8.3 Secondary Outcome measures

The secondary outcome measures are:

Besides our main outcome (Paranoia Scale), we have identified 5 key secondary outcomes (two measures of the target mechanism: SRT, SST; three measures of clinical symptoms: GPTS, HADS, PWQ) upon which we will perform hypothesis testing:

- Paranoid beliefs (SRT, SST)
- Paranoid beliefs measured using an 8-item version of the *SRT* and the SST
- Clinical symptoms assessed by *Paranoia Scale* and *R-GPTS*
- State mood change measured pre-post each session using visual analogue scales (VAS: anxious, sad, paranoid, friendly).

See study protocol for details on the secondary outcomes.

8.4 Follow-up assessments

Follow-up assessments will include the full battery of outcome measures and will be given at 6-, 12-, 18-, and 24-weeks (see Protocol for details).

8.5 Interim assessments

The design includes every arm receiving weekly post-session 'interim' assessments for 12 weeks. Sessional assessments comprise a small subset of the main outcome measures as follows:

- Paranoid beliefs measured using an 8-item version of the *SRT* and the SST
- Clinical symptoms assessed by *Paranoia Scale* and *R-GPTS*
- State mood change measured pre-post each session using visual analogue scales (VAS; anxious, sad, paranoid, friendly)

9 Principal Research objectives

9.1 Estimand

The estimand will be based on the Intention-to-treat (ITT) principle and the treatment policy strategy, as defined by the International Council for Harmonisation (ICH E19 2019). Under the ITT principle, all patients will be included in the analysis and they will be analysed as part of the treatment group to which they were originally assigned to, irrespective of treatment cross over or treatment discontinuation. The treatment policy strategy has an almost identical meaning to the ITT principle in terms of being pragmatic, since under this strategy we are still comparing treatments under the conditions in which they would be used in practice. The main difference is that the ITT principle focuses more on which subjects should be included in the analyses, whereas the treatment policy strategy focuses on which data should be included in the anaysis in case of intercurrent events (intercurrent events are defined in Section 10.1).

For this trial, the ITT/treatment policy estimand is defined as the average treatment effect (reduction of Paranoid symptoms measured using the Paranoia Scale) between 6 sessions and 12 sessions versus control at 24-week follow-up, regardless of adherence to the allocated study treatment (or other intercurrent events that could occur) for all randomized individuals. Subjects allocated to a treatment arm will be followed up, assessed and analysed as members of that arm irrespective of their compliance to the planned course of treatment. Under this estimand, we will carry on collecting outcome data after treatment discontinuation and treat missing data after treatment discontinuation as though they had been observed (ICH E19 2019).

9.2 Primary objective

To evaluate 6 and 12 sessions of STOP mobile App's plus TAU effectiveness (superiority) in the reduction of paranoid symptoms measured by the Paranoia Scale at 24-weeks post-randomisation compared to control app sessions plus TAU.

9.3 Secondary Objectives

To investigate the effectiveness of 6 and 12 sessions STOP mobile App plus TAU compared to 12 control app sessions plus TAU, in people who experience clinical levels of paranoia, based on:

1. the reduction of paranoid beliefs measured by a) the Similarity Rating test (SRT; Mathews & Mackintosh (2000)) and b) the Scrambled Sentences task (SST; Rude et al. (2003)) at baseline, 6, 12, 18, and 24 weeks post randomisation.
2. the reduction of clinical symptoms measured at baseline, 6, 12, 18 and 24 weeks post-randomisation, using:
    2.1. Green Paranoid Thoughts Scale Revised (R-GPTS; Freeman, D., et al., 2021)
    2.2. Hospital Anxiety and Depression Scale (HADS; Zigmond & Snaith (1983))
    2.3. Paranoia Worries Questionnaire (PWQ; Freeman, D., et al., 2020)
    2.4. Positive and Negative Symptom Schedule (PANSS; Kay, Fiszbein & Opler (1987)) - item 6 only
3. Recovery, measured at baseline, 6, 12, 18 and 24 weeks post-randomisation, using:
    3.1. the Warwick-Edinburgh Mental Wellbeing Scale (WEMWBS; Tennant, R., et al. (2007))
    3.2. Questionnaire about the Process of Recovery (QPR; Neil, S.T., et al. (2009))
    3.3. Recovering Quality of Life (ReQoL; Keetharuth AD, et al. (2018)) questionnaire

Inference will only be performed on SRT and SST and the three clinical symptom measures GPTS, HADS and PWQ, see 12.6 Method for handling multiple comparisons

User satisfaction with the mobile app is measured with the Ben-Zeev et al., (2014) instrument given once only at the end of each arm's final session.
Visual Analogue Scales (VAS: anxious, sad, paranoid, friendly) will be used pre and post each session to monitor transient fluctuations in mood across individual sessions.

9.4 Exploratory Objectives

The study was designed to include 12 weekly post-session interim assessments to model the relationship between the duration of the therapy and paranoia symptoms in a dose-response curve analysis of the 6-session CBM-pa therapy plus TAU and the 12-sessions STOP mobile App therapy plus TAU compared with the 12-session text reading control plus tau.

10 Intercurrent events and Treatment adherence, protocol violations, withdrawal from treatment and trial, and early trial stopping

10.1 Intercurrent Events

Intercurrent events are events that occur after randomization and which either preclude the observation of the outcome or affect its measurement and/or interpretation (ICH E19 2019, Kahan et al. 2021). Potential intercurrent events related to the disease or the specific intervention are:

- Non-adherence to treatment (treatment discontinuation, missed sessions or not starting treatment): expected to be independent of treatment arm due to blinding
- Use of non-trial treatments: Expected to be independent of treatment arm due to blinding and not problematic for this trial
- Rescue treatment and/or discontinuation due to an SAE
- Treatment switching: Not possible due to the design of the app
- Death: unlikely to be related to the trial

Handling of intercurrent events:

As primary and most secondary outcomes are collected through the app, we will collect and include all data from all cases regardless of whether intercurrent events related or unrelated to the treatment or disease (e.g. moving away) happened or not. Death is considered an unlikely event in this study.

10.1.1 Adherence

Adherence to the intervention is defined as outlined in the table below.

Table 1. Definitions of Adherence

		Acceptable Values	Target Values
Individual sessions	To 'complete' an individual session participant must….	enter responses on at least 20 out of 40 trials (50%)	enter responses on at least 30 out of 40 trials (75%)
Intervention overall	To 'complete' the intervention as a whole participants must….	Complete ¼ of all sessions scheduled to date	Complete ½ of all sessions scheduled to date
Number of participants	**Proportion of sample to be adherent in order to meet funder milestones**	**50%**	**75%**

To facilitate adherence, STOP will include proven gamification techniques including progress tracking and reward systems (see Protocol for details).

As stated in the Protocol, the following boundaries will apply to the delivery of interventions and assessments, all of which are incorporated within the programmed functionality of the app and are therefore implemented automatically:

Interventions

The intervention will comprise 12 predefined, consecutive calendar weeks for each participant, starting on the first Saturday after date of randomisation.

- Each calendar week runs from 00.00 Saturday – 23.59 Friday. Saturday is chosen as the start of the week to allow maximum time for users to miss and reschedule their sessions.

- Participants and researchers will schedule the first session together, which then prepopulates the remaining sessions, all 7 days apart.
- Researchers will ensure that there are at least two days between randomisation and the first scheduled session (a technical requirement of the app software system).
- Subsequent sessions can be individually adjusted (e.g. to accommodate other commitments) as needed within the following allowable parameters:
    - Each session must occur at sometime within its 7-day allocated calendar week (this prevents problematic variance in intensity of treatment if using a 7 days +/- 2' condition where one ppt can always go for 5 days apart – giving a 60 day tx window - and another 9 days apart – giving a 108 day tx window).
    - Each session must be at least 2 full days apart (thus Friday of week 1 can be followed by either Monday of week 2, or the last day of week 2, Sunday. Session 3 would then have to fall, in the first case, between Monday and Sunday of week 3 and in the second case between Wednesday and Sunday of wk 3. Sessions can therefore be between 2 and 12 days apart)
- Participants are encouraged to schedule all sessions early in the calendar week (i.e., close to Saturday) to allow maximum time for rescheduling if needed, thereby reducing the risk of missed sessions.
- Participants are asked to plan to complete each session in a single sitting. (subject to piloting) If unavoidably interrupted they should complete within the same day. Sessions not completed the same day will trigger an alert to researchers the next day who can follow up case by case.
- Sessions remain open for completion up until 2 days prior to the next scheduled session (i.e. consistent with the minimum allowable time between consecutive sessions)

Assessments

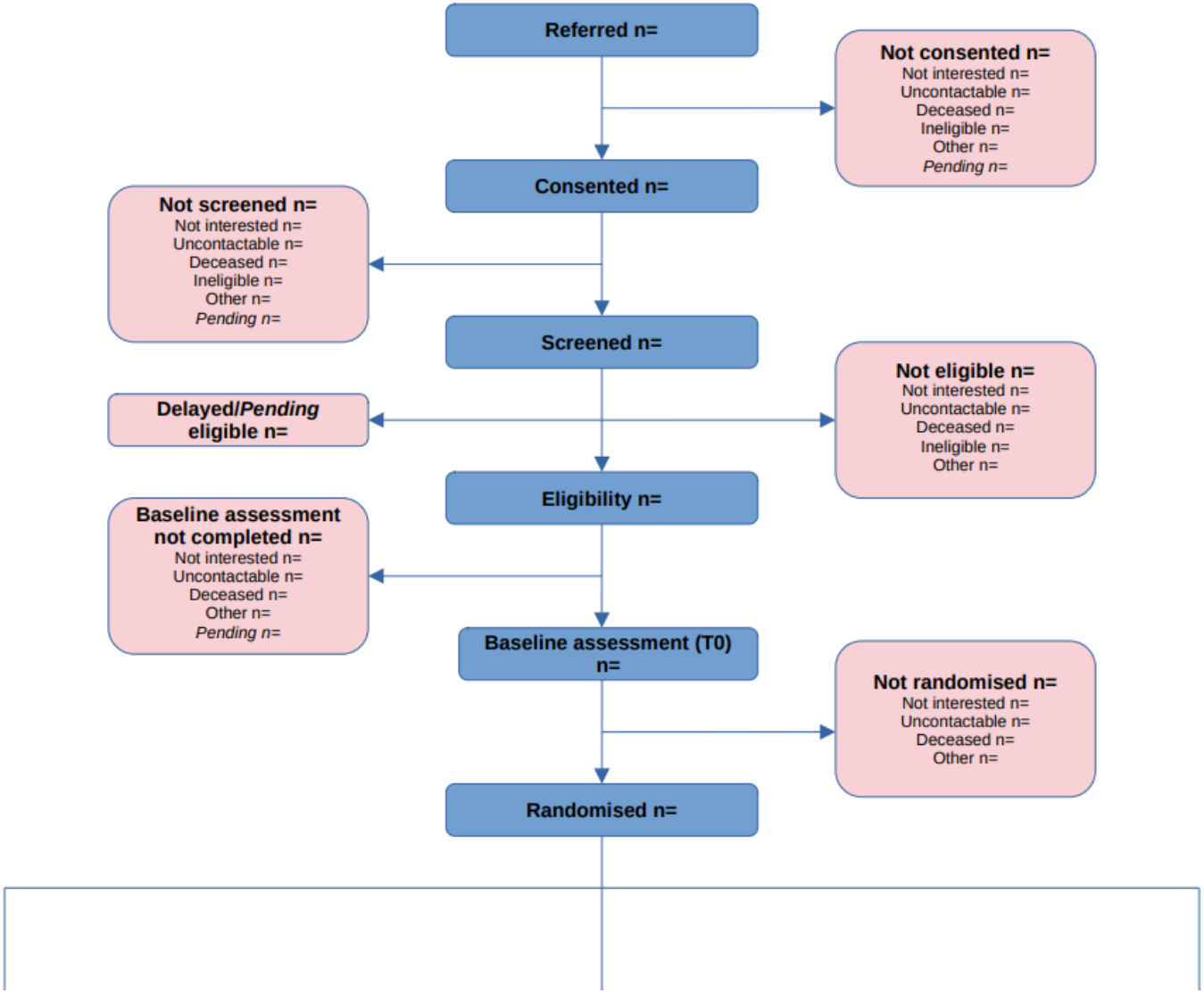
Follow up assessment dates are calculated as 6, 12, 18, 24 weeks post randomisation date.

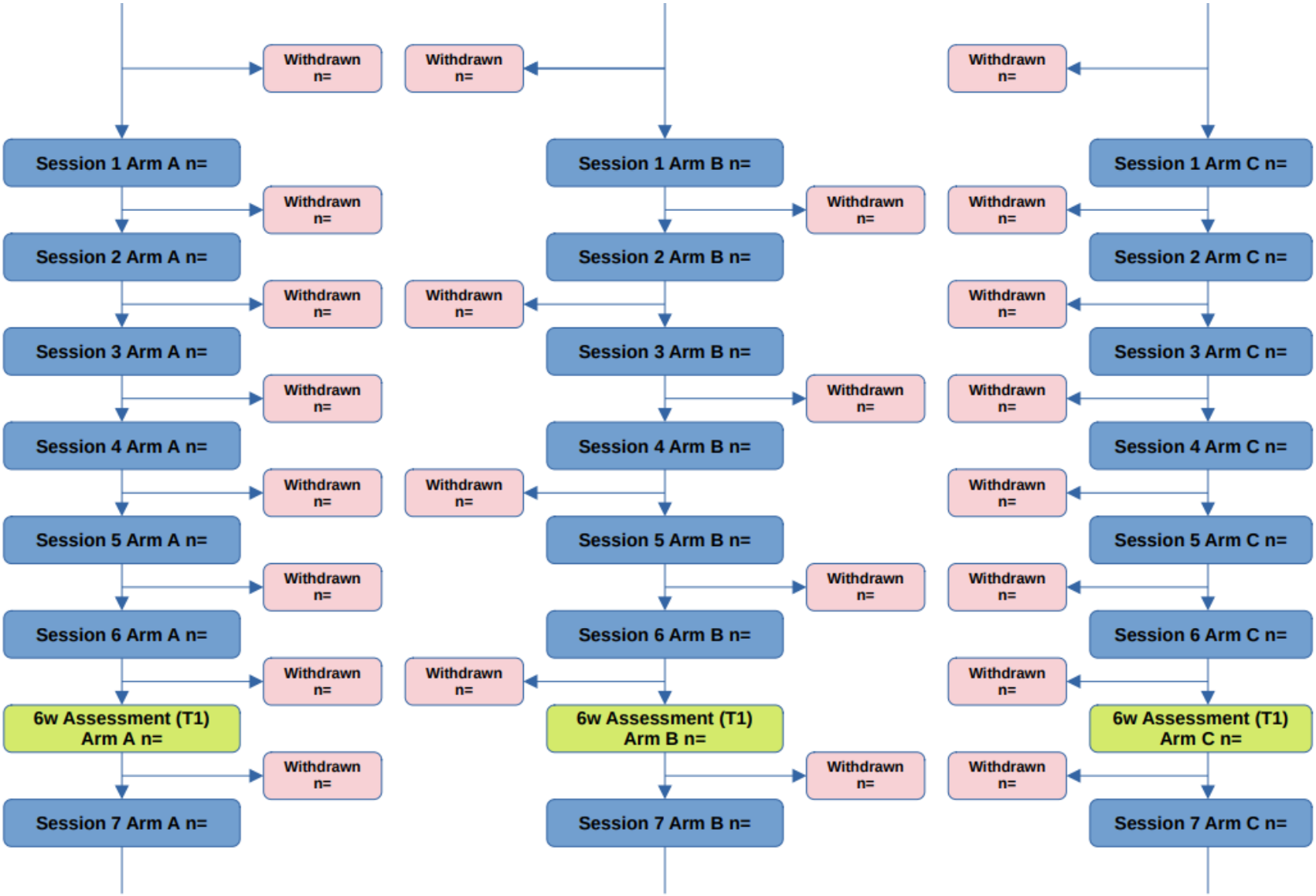
A window of +/- 3 days either side is allowed, with the exception of the 6 week assessment, which is automatically scheduled at the mid-point between session 6 and session 7.
Researcher contact is required at each follow up assessment to administer PANSS 6 and maintain contact.

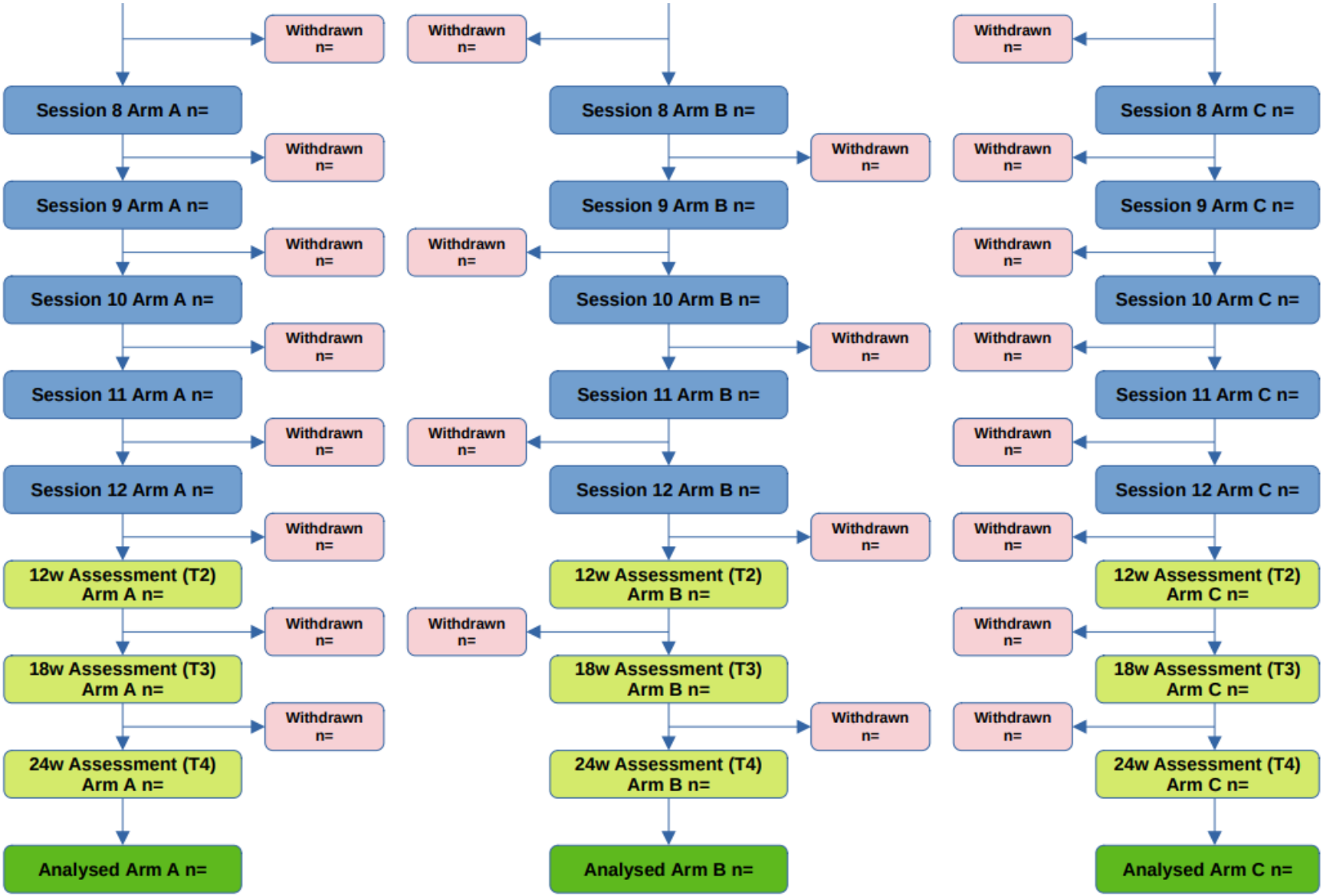
All participants receive all 12 interim assessments (those randomised to the 6 session arm will receive assessments only in weeks 7-12) . This serves 2 purposes: a) provides control data for the secondary dose response analysis and b) prevents unblinding that could result from scheduling a different number of sessions.

Presentation of adherence

Adherence data will be presented to the Data Monitoring Committee (DMC) and recorded in the Consort flow chart below:


Referred n=
Not consented n=
Not interested n=
Uncontactable n=
Deceased n=
Ineligible n=
Other n=
Pending n=
Consented n=
Not screened n=
Not interested n=
Uncontactable n=
Deceased n=
Ineligible n=
Other n=
Pending n=
Screened n=
Not eligible n=
Not interested n=
Uncontactable n=
Deceased n=
Ineligible n=
Other n=
Delayed/*Pending* eligible n=
Eligibility n=
Baseline assessment not completed n=
Not interested n=
Uncontactable n=
Deceased n=
Other n=
Pending n=
Baseline assessment (T0) n=
Not randomised n=
Not interested n=
Uncontactable n=
Deceased n=
Other n=
Randomised n=


Withdrawn n=
Session 1 Arm A n=
Session 2 Arm A n=
Session 3 Arm A n=
Session 4 Arm A n=
Session 5 Arm A n=
Session 6 Arm A n=
6w Assessment (T1) Arm A n=
Session 7 Arm A n=
Session 1 Arm B n=
Session 2 Arm B n=
Session 3 Arm B n=
Session 4 Arm B n=
Session 5 Arm B n=
Session 6 Arm B n=
6w Assessment (T1) Arm B n=
Session 7 Arm B n=
Session 1 Arm C n=
Session 2 Arm C n=
Session 3 Arm C n=
Session 4 Arm C n=
Session 5 Arm C n=
Session 6 Arm C n=
6w Assessment (T1) Arm C n=
Session 7 Arm C n=


Withdrawn n=
Session 8 Arm A n=
Session 9 Arm A n=
Session 10 Arm A n=
Session 11 Arm A n=
Session 12 Arm A n=
12w Assessment (T2) Arm A n=
18w Assessment (T3) Arm A n=
24w Assessment (T4) Arm A n=
Analysed Arm A n=
Session 8 Arm B n=
Session 9 Arm B n=
Session 10 Arm B n=
Session 11 Arm B n=
Session 12 Arm B n=
12w Assessment (T2) Arm B n=
18w Assessment (T3) Arm B n=
24w Assessment (T4) Arm B n=
Analysed Arm B n=
Session 8 Arm C n=
Session 9 Arm C n=
Session 10 Arm C n=
Session 11 Arm C n=
Session 12 Arm C n=
12w Assessment (T2) Arm C n=
18w Assessment (T3) Arm C n=
24w Assessment (T4) Arm C n=
Analysed Arm C n=

10.2 Protocol violations

Protocol violations are any occurrence that results in a deviation from the procedures outlined in the study protocol, whether through human error or technical fault. All protocol violations will be recorded in the following table:

Date occurred	**Date reported**	**Participant MACRO Pin (s)**	**Category** 1. Unblinding 2. Enrolment 3. Randomisation 4. Study contact (recorded in App; reported to DMC as adherence data) 5. Data collection 6. Safety reporting 7. Other	**Description**	**Results from the action of** 1. Staff 2. Participant	**Action taken**	**Signed by CI (Y/N)**

10.3 Withdrawal from treatment and trial

Participants choosing to take part will follow the procedures as predefined for their allocated arm (see Protocol). It will be made clear to each participant at the time of informed consent that they may withdraw from the trial at any time without having to give a reason and without this impacting on their usual clinical care in any way. In addition, state mood will be measured before and after each session using the visual analogue scales anxious, sad, paranoid, friendly. State mood normally fluctuates widely on a day to day and moment to moment basis, however a consistent pattern over time can indicate more enduring effects that should be investigated further. Therefore, an automatic alert will be sent to researchers if a participant's mood worsens across the session on 3 consecutive sessions. Following an alert, a researcher will contact the participant to make a further assessment, including completing the Adverse Event checklist and escalate within the research team if appropriate, and, if necessary, procedures for notifying an adverse event or serious adverse event will be triggered. The number and proportion withdrawing from the trial (i.e., actively state they are unwilling to

provide any further research data), and the reasons for withdrawal will be summarised by intervention group and overall. Withdrawal will be presented to DMC and recorded in the Consort flow-chart shown in section 10.1. The flow-chart will capture the reasons and the timing of the withdrawal.

10.4 Early trial stopping

The trial may be prematurely discontinued by the MRC should key Milestones remain unmet or based on new safety information or for other reasons. The DMC/TSC can also independently recommend discontinuing the study. If the study is prematurely discontinued, active participants will be informed and withdrawn and no further participant data will be collected.

10.5 Loss to follow-up and other missing data

The number and proportion of participants missing each primary and secondary outcome variable will be summarised by intervention group and overall, at each time point. Missing data in the database will be presented in compliance with CONSORT diagram (Boutron et al., 2008).

11 Adverse event reporting

AEs are defined by the Health Research Authority (HRA) as any untoward medical occurrence, unintended disease or injury, or untoward clinical signs in trial participants, whether or not related to the intervention which require additional support or input from health professionals. The study will follow the guidance given by MEDDEV 2.7/3 Rev 3 Clinical Investigations: Serious Adverse Event Reporting (https://ec.europa.eu/docsroom/documents/16477/attachments/1/translations) [61] and by the HRA (https://www.hra.nhs.uk/approvals-amendments/managing-your-approval/safety-reporting/)

11.1 Urgent reporting

If an urgent safety measure is required to protect research participants against any immediate hazard to their health or safety, this will be implemented without prior authorisation from a regulatory body. Please see study protocol for details

11.2 Regular Reporting of serious adverse and adverse events

All adverse events (AE), and serious adverse events (SAE) will be summarised as the number of events and the number of people having events by intervention group and overall and reported to each meeting of the DMC, who report to the Trial Management Committee.

These aggregated statistical reports will be generated from the study's MACRO database via the Adverse Event Log or at any time at the request of the DMC Chair. Full reporting procedures are outlined in the study's Adverse Event SOP. These include an annual report to the REC and immediate report to the MHRA and chair of the DMC of every individual serious adverse event (related or not). The DMC will be responsible for investigating further if there are any concerns about unexpectedly high rates of SAEs, which may include being unblinded as to trial condition or seeking further data on adverse events and will advise the TSC on any ethical or safety reasons why the trial should be prematurely ended. The Funder will immediately be notified on receipt of any information that raises material concerns about safety or efficacy, and of any recommendations from the DMC to end the trial.

12 Data analysis plan

12.1 Baseline comparability of randomised groups

The description of variables measured at baseline is in the study protocol. All baseline variables listed in Section 8.1 will be described overall and by intervention group. Frequencies and proportions will describe categorical variables, while numerical variables will be described using mean and standard deviation (SD) if normally distributed and median and interquartile range (IQR) if not normally distributed. No statistical testing of the baseline differences between randomised groups will be done.

12.2 Descriptive statistics for outcome measures

The primary and secondary outcomes will be summarised at baseline, six weeks, twelve weeks and 24 weeks post-randomisation by intervention group and overall. Mean and SD or medians and IQR will be used to summarise normally distributed and non-normally distributed variables, respectively.

12.3 Inferential analysis

The analyses of effectiveness outlined in this strategy will be pragmatic, and will utilise all available follow-up data from all randomised participants

All analysis will follow the intention to treat principle (Fergusson, Aaron et al. 2002). Group difference in mean estimates and associated 95% confidence intervals between 1) the 6-session treatment arm and control, and 2) between the 12-session treatment arm and control will be reported. These will be estimated using models that account for the repeated measurements of the participants, as described below. The main statistical analyses aim to estimate the group mean differences at the 6th week (end of the 6-session treatment), the 18th week, and the 24th week post-randomization (6 weeks and 12 weeks post-treatment observation time – see Section 2).

Adjustment for covariates

In this study, patients are recruited from two sites (London and Bath) and randomization is stratified by site and sex at birth and hence, the analyses models will always include study site and participant sex at birth. Post-randomization measurements are taken at 6, 12, 18 and 24 weeks; we expect an increasing drop-out rate as treatment progresses and as a result, we will analyse all four time-points simultaneously, under (restricted) maximum likelihood, to reduce potential biases and to maximize power.

12.4 Analysis of primary outcome

As previously stated, the primary outcome is Paranoia Scale (PS; Fenigstein & Vanable) and the primary objective the assessment of the reduction of paranoid symptoms in terms

of mean PS estimates at 24 weeks post-randomization. PS ranges from 20 to 100 (see Protocol) and will be treated as continuous. Data from our feasibility study showed that the primary outcome was approximately Normally distributed ranging between 20 and 80 points. To evaluate the treatment's effect of a 6- and a 12-session of the STOP mobile App plus TAU compared to a 12-session text reading control plus TAU, a linear *mixed-model repeated measures* (mmrm) approach (Mallinckrodt et al. (2001)) will be fitted using restricted maximum likelihood (reml). The model will allow a separate mean paranoia scale (PS) parameter at each assessment time in each treatment group and an unstructured covariance matrix of the response-level residuals.

Under the repeated measures setup, we will treat baseline outcome and dummy variables for the randomisation factors (study site, sex at birth) as independent variables with a design matrix that allow no treatment difference at baseline (though a different residual variance at follow-up with non-zero covariances over time and across treatment groups). Therefore, *treatment randomisation group* (with levels denoting control, 6-sessions, and 12-sessions treatment), *time* (with levels denoting weeks 6, 12, 18 and 24 weeks), *baseline PS score*, *treatment by time interaction*, *baseline PS score by time interaction*, *study site*, and sex at birth will constitute the fixed part of the model. Continuous baseline covariates will be centred at their means. This setup allows us to estimate the treatment effect at 24 weeks post-randomisation and lets follow-up data be treated as ignorable under a missing-at-random assumption, while the model accounts for the possible imbalance due to random sampling in baseline measurement of the outcome variable to control for pre-treatment differences.

The mmrm model will be fitted using Stata's mixed command as:

```
mixed PS i.time i.treatment c.baseline_PS i.time#i.treatment
i.time#c.baseline_PS i.study_site i.sex at birth || participant_id:, nocons
residuals(unstructured, t(time) by(treatment)) reml
```

where:

PS: continuous dependent variable denoting mean paranoid symptoms, measured at weeks 6, 12, 18 and 24.

i.time: categorical independent variable indicating assessment time, coded as: 0 (6 weeks), 1 (12 weeks), 2 (18 weeks), 3 (24 weeks).

i.*treatment*: categorical independent variable denoting the treatment arm, coded as: 0 (control), 1 (6-week treatment), 2 (12-week treatment).

c.*baseline*_PS: continuous independent variable denoting mean paranoid symptoms at baseline.

i.*time*#i.*treatment*: interaction between categorical time and categorical treatment.

i.*time#c.baseline_PS*: interaction between categorical time and continuous baseline paranoia symptoms.

i.study_site: categorical independent variable indicating study site, coded as: 0 (London), 1 (Bath).

i.sex at birth: categorical independent variable indicating sex at birth, coded as: 0 (females), 1 (males).
participant_id: the participant's ID.

|| participant_id:, nocons: instructs Stata that the data are clustered by participant. The model allows for response-level variability only and assumes that participant-level variability has been explained away by randomisation; the option nocons stops Stata fitting a random intercept for each participant.

residuals(unstructured, t(time) by(treatment)): estimates a residual covariance matrix based on the data at hand (unstructured) for each treatment arm separately (by(treatment)): an unstructured matrix does not impose any particular form of correlation between the unexplained part of the responses taken at two different time points, and estimates different residual variances for each time point.

reml: instructs Stata to fit the model using Restricted Maximum Likelihood (Harville, 1977).

The mathematical form of the model is:

$$y_{ijk} = c + \sum_{i=1}^{3} \beta_{Time_i} t_i + \sum_{k=1}^{2} \beta_{Trt_k} trt_k + \beta_1 y_{0j} + \sum_{k=1}^{2} \left(\sum_{i=1}^{3} \beta_{TimeTrt_{ik}} t_i \right) trt_k + \left(\sum_{i=1}^{3} \beta_{TimeBase_i} t_i \right) y_{oj} + \beta_2 s_j + \beta_3 g_j + \varepsilon_{ijk}$$

where:

y_{ijk}: vector of paranoia scale values taken at time i from participant j in treatment arm k, with i = 0 (6 weeks), 1 (12 weeks), 2 (18 weeks), 3 (24 weeks), j = 1,….,N, and k = 0 (control), 1 (6-week STOP), 2 (12-week STOP).

c: the overall intercept.

t_i: three dummy variables that correspond to week 12, week 18, and week 24.

trt_k: two dummy variables, one for the 6-session STOP treatment and one for the 12-session STOP treatment.

y_{0j}: the baseline score on the paranoia scale for participant j centred at its mean.

s: a dummy variable for the study site Bath.

g: a dummy variable for the sex at birth male.

$\beta's$: coefficients of the dummy and the continuous variables.

ε_{ijk}: response errors for time i from participant j in treatment arm k.

Mean differences of treatments with 95% confidence intervals at follow-up will be calculated via Stata's `lincom` command, which enables the estimation of linear combinations of model parameters and their degree of uncertainty. In addition, standardised effect sizes (Cohen's d calculated as estimated mean treatment difference divided by the standard deviation of baseline outcome) with bootstrapped 95% confidence intervals will be calculated. Bootstrap will be performed using Stata's bootstrap command with the option cluster() to resample study participants, not observations. In order to ensure that the resampled participants are uniquely identified in each bootstrapped dataset, we will also add the option idcluster(). This creates, at each replication, a new variable that contains a unique identifier for all participants.

12.4.1 Model assumption checks

The linear mixed effects models assume normally distributed residuals. This will be checked when describing the data. Outcome residuals for each time-point will be plotted using a Q-Q plot to check for Normality and the existence of outliers. If violations of the assumptions are observed, bootstrapped 95% confidence intervals will be calculated using the bootstrap procedure, described above.

12.5 Analysis of secondary outcomes

The analyses of effectiveness outlined in this section will be pragmatic, based on intention-to-treat and will utilise all available follow-up data from all randomised participants.

For the secondary outcomes, we will use the same model to evaluate the effect of the STOP intervention, provided that the outcome values can be assumed to be Normally distributed. Where this is not the case, we will make use of generalised random effects models with robust standard errors, under the missing-at-random assumption about the data. For binary outcomes we will fit random effects logistic models with Stata's melogit command and for ordered outcomes random effects ordinal logistic models with Stata's meologit command. The syntax of these model is analogous to that for the mixed model except that we also specify the number of quadrature points using the `intpoints()` option. As an example of an ordinal outcome, the Stata code will look like:

```
meologit outcome i.time i.treatment c.baseline_PS i.time#i.treatment
i.time#c.baseline_PS i.study_site i.sex at birth i.paranoia_severity ||
participant_id:, intpoints(30)
```

Logistic and ordinal regression will provide only subject-specific estimates. In addition to the subject-specific estimates, we will also provide approximate population estimates (PE). According to Carpenter and Kenward (1997), population estimates for treatment effects can then be obtained as:

$$\beta^{PE}_{TimeTrt_{ik}} = \frac{\beta_{TimeTrt_{ik}}}{\sqrt{1 + 0.34584\sigma^2_u}}$$

where:

$\beta_{TimeTrt_{ik}}$: the effect of treatment k at time i from the random effects model.

σ^2_u: the variance of the random intercepts.

95% confidence intervals for $\beta^{PE}_{TimeTrt_{ik}}$ will be calculated via bootstrap, described above.

12.6 Level of significance and methods for handling multiple comparisons

The alpha error level for our primary hypothesis is set to 0.05. We control the type 1 error for our two planned and orthogonal contrasts (6 sessions against control and 12 sessions against control at 24 weeks follow-up) by using the LSD test procedure, which does not require a correction of the familywise error of pairwise comparisons of three groups if the main effect is significant. No further correction will be made for multiple comparisons involving the primary outcome.

As well as our main outcome (Paranoia Scale), we have identified 5 key secondary outcomes (two measures of the target mechanism: SRT, SST; three measures of clinical symptoms: GPTS, HADS, PWQ) for which we will perform hypothesis testing. To account for multiple testing in the secondary outcomes, significance will be reported at the 1% level, which reduces considerably the chance of false positive results. Tests with a p-value between 1% and 5% will be regarded as trends.

For the remaining variables, we will provide statistics and estimated treatment effects as purely explorative results and will not perform formal statistical tests. Standard errors will be provided for the reader to make their own formal assessment if desired (Senn, 2017).

12.7 Missing data

If there is complete follow-up, an estimand based on treatment policy can be estimated with little assumptions. However, despite efforts to collect data on all randomized participants, invariably there will be some missing data. We do not expect missing data at baseline. Missingness at the baseline of the main clinical outcome variables is not possible due to the recording method in the app.

12.7.1 Minimizing attrition at follow-up

We will try to minimize attrition at follow-up by reminding participants to participate in the trial. We will send reminders the day before their scheduled follow-up measurements by email and by using automatically generated reminders on their mobile phones. Also, researchers will be present at follow-up appointments and will be aware of the need to keep data loss to a minimum.

However, some subjects may discontinue from the trial and others may miss one or two of the three follow-up visits. Reasons for these events will be recorded, see MACRO Withdrawal form. The missing data will be addressed in the statistical analyses using principled methods such as Maximum Likelihood, which can provide an unbiased and precise estimate of the ITT estimand the presence of missing data at random.

When missing data exist, analyses rely on assumptions about the behaviour of the participants after dropping out. A robust estimate can only be made if the assumptions that need to be made are justifiable and plausible. Therefore, assumptions need to be explicitly stated. In practice, however, assumptions are hard to be proven and as a result, sensitivity analyses will be performed to explore the robustness of the inference to these assumptions.

12.8 Reporting of missingness

We will follow CONSORT guidelines of reporting the reasons why patients were lost of follow-up and hence why outcomes are missing. We do not expect a differential loss of participant counts between arms because participants will not be aware if they are randomized into a treatment or control group. Therefore, bias due to missingness should be similarly distributed among treatment arms. This assumption will be assessed by comparing the number of participants between arms. To identify the potential for bias due to missing data, we will present a table comparing the distribution of baseline data by treatment arm for patients with observed follow-up data and for patients with missing follow-up data separately for each study arm (Schultz et al 2010). Missing data will be described both as a proportion in each variable separately as well as using a graph that ranks variables from most complete to least complete. This will be done for all participants together as well as for each arm separately. Due to the data collection via app, we expect that most clinical outcome data are either all collected for a participant or all are missing. Given the amount of missing data we encountered in the feasibility study, we anticipate a high degree of completeness.

12.8.1 Missing data in baseline variables

If the proportion of missing data in baseline variables is small (<=2%) we will fill the missing data back in with the mean of the baseline variable (White & Thompson, 2005). If the proportion of missing data in the baseline variables is large, we will consider using multiple imputation (Rubin (1987)) by chained equations, which can accommodate the imputation of missing data in baseline variables as well as in outcome measures at the same time. Results from this process will be contrasted with results from the complete case data analyses.

12.8.2 Missing data in the outcome

The described mixed model will be estimated via restricted maximum likelihood that is valid under the missing at random (MAR) assumption and allows for the inclusion of all patients with at least one follow-up observation. This model uses an unstructured covariance structure and is appealing due to its robustness against model misspecifications and its unbiasedness when data are missing completely at random or missing at random, assuming that subjects who withdraw from the study at any time point would have continued just like their peers in

the same arm who have the same observed outcome data (Carpenter et al. 2013). This last statement points towards the MAR assumption which requires that all predictors of the missingness mechanism be included in the models to maximise the likelihood that, conditional on these predictors, the outcome is missing at random. For our main statistical analyses, we will be assuming that the data are MAR, but we will also perform the following sensitivity analyses to explore the robustness of conclusions to departures from MAR.

1) Including predictors of missingness in the model

To make the assumption of missing at random more plausible, we have prespecified baseline variables that are assumed to be potential predictors both of the incomplete variables and of the probability of a value being observed. These prespecified variables are age, gender, ethnicity, education level, employment, living arrangement, relationship status, IQ (WTAR), self-reported dyslexia, age of onset of distressing paranoia, Paranoid beliefs, Paranoia Scale, and HADS. For the primary analysis we will exclude Paranoia Scale from this list since it will be used as a dependent variable.

We will perform logistic regressions (with missing at follow-up (yes/no) as dependent variable) to identify key variables among the prespecified variables that are predictive of missingness in the analyses following the guidelines of Carpenter and Smuk (2021).

To assess whether missing outcome data are predicted by baseline variables, we will firstly construct binary (0/1) indicator variables per outcome, capturing whether any of the outcomes are missing their 6-, 12- 18- or 24-week value, respectively. These indicators will constitute the dependent variable in logistic regression models that include intervention group, PS score at baseline, sex at birth and study site as independent variables. We call these models “core missingness” models. Then, we will examine whether several baseline variables (see next paragraph) correlated with the outcome of interest, can predict missingness in the outcome when added to the core model, by following two steps. In the first step, we will add the baseline variables in the core model, one at a time. They will be considered to predict missingness if there is a significant relationship at a 5% level. In the second step, the significant variables identified in step one, will enter a new model all at the same time. The new model will be a logistic model with the same binary indicator of “missingness” used in step one. The statistically significant variables from this will constitute the predictors of missingness and will

be incorporated in the models for the analysis of the primary or secondary outcomes as part of a sensitivity analysis.

The baseline variables we will examine are: age, gender, ethnicity, education level, employment, living arrangement, relationship status, IQ (WTAR), self-reported dyslexia, age of onset of distressing paranoia, Paranoid beliefs, Paranoia Scale, and HADS. For the primary analysis we will exclude Paranoia Scale from this list since it will be used as a dependent variable. All significance levels will be at the 5% level. We will report how sensitive our results are to the inclusion of these variables and any changes to estimated treatment effects.

2) Sensitivity of results due to missing data by assuming missing not at random (for primary analysis)

Analysis of data where the outcome is incomplete always requires untestable assumptions about the missing data commonly that the data are missing at random. The estimand would be inappropriate if a large proportion of participants leave the study due to intercurrent events and are not available for the final endpoint measurement. In this case, we assume that the data are potentially not missing at random (MNAR). To address this the primary analyses will be conducted using a pattern mixture with multiple imputations model (Carpenter et al 2016). This approach addresses the MNAR problem that arises when the implied distribution of the data after drop-out cannot be justified under the postulated model. As such, a different distribution can be specified for a different pattern of missingness of (groups of) participants that can reflect a specific assumption appropriate to their treatment arm, drop-out time and possibly other relevant information. We will perform sensitivity analyses to explore the effect of departures (varied over a plausible range) from the assumption of missing at random made in the main analysis following the steps recommended by White et al (2005). We will set up a multiple imputation model consistent with the primary substantive analysis model. We will use multiple imputations to impute missing values under a MAR assumption and modify the MAR imputed data to reflect a plausible range of patterns related to the effect of non-compliance followed by drop-out following the guidelines of Carpenter et al 2013).

We will explore three types of information-anchored sensitivity analysis:

1. One approach acknowledges that it is possible that drop-out is caused by the app itself and not the treatment (we assume that participants do not know if they are in treatment or control conditions) and the app itself may have a negative influence. To examine this, a delta-adjusted approach will be performed, where the MAR imputed values are modified to reflect the assumed MNAR mechanism (Carpenter and Kenward 2013). Participants who discontinue the treatment become worse by some fixed amount compared with participants who continue the trial. After imputing the data under MAR, we will increase the imputed values by 2%, 5%, 7%,... of the treatment group means at the main time point, until we reach the 'tipping point' where the treatment effect is no longer significant. A-priori, we believe that a 5% difference is plausible so that if the test for treatment effect remains significant, we will report the results are 'robust to plausible a-priori agreed departure from MAR.'
2. Another approach is the copy reference-based pattern mixture model which assumes that the post-dropout response profile of participants who discontinue the randomized treatment will be like that of patients on the TAU arm. We will use 'Copy Reference' as a more extreme sensitivity analysis (Carpenter et al 2016), whereas reference we will choose the control arm. Under Copy Reference when participants drop out of the study the means and the covariance matrix of their response distribution, both before and after drop-out, are replaced entirely with those from the reference arm. Hence, if the reference group is assumed to be the control arm, then this assumption mimics the case where those dropping out do not respond to treatment. For participants in the reference arm, their imputation model is that of the MAR assumption.
3. A third approach is to assume that the app and the treatments have additive negative influences on the participants which result in drop-out. Here we would assume differential worsening of the clinical outcome by the different arms.

3.) Terminal Events

Death is classified as an intercurrent event, but the treatment policy strategy cannot be implemented for intercurrent events that are terminal events, since values for the variable after the intercurrent event do not exist. In the unlikely event of a death, we will, therefore,

expand our sensitivity analysis to include a scenario where the terminal event would not occur (i.e., the life of the participant was saved) and add the value that the clinical outcome would have taken in the defined hypothetical scenario (i.e., getting worse). This scenario can be added to any of the three pattern mixture models described above, as additional instances (patterns) that assume extreme worsening of the outcome. Finally, if death is related to the treatment and/or the app, our treatment policy strategy will not be relevant anymore as this would be reported as a serious adverse event that would prompt external investigations into the continuation of the trial. Any deaths confirmed as unrelated to the treatment and/or the app will be assumed to be MAR cases, but MNAR sensitivity analyses will still be performed. All scenarios will be prespecified in the analyses and the range of shifts of imputed values will be guided by content experts. Additional scenarios will be performed on the request of the DMC committee. We will then analyse the data of the different models as per our primary analyses following the standard recommendation of multiple imputation guidelines to obtain a single MNAR estimate and standard errors and report any changes in the estimated treatment effects.

12.8.3 Missing data in scales and subscales

This paragraph is not relevant for the primary outcome variable but only for some of the secondary outcome and baseline variables which are based on questionnaires or similar item-based measurement scales. The planned strategy for handling missing data for the item and scales will depend on the amount of missing data observed and the planned analyses for the outcomes.

Missing item data will be imputed using pro-rating that is, by replacing the missing item score by the mean of the observed items if less than 20% of the items scores are missing. Items within each scale are indicators of a specific concept and as a result assumed to be closely and positively correlated and therefore regarded as a particularly applicable technique [Downet and King 1998, Roth et al 1999), (8, 9]. Simulation studies have shown that pro-rating (or case mean substitution) is a robust method when data are missing on less than 20% of items in both random and systematic patterns [Roth et al 1999].

Pro-rating is implemented across items within a scale, or subscale, for each assessment and participant. In the unlikely event that more than 20% missing items on an item variable data is collected we will treat it as a missing data point and proceed as in Section 12.6.2.

To ensure the same strategy is followed across all scales reported in the principal paper(s), any guidance given by authors of validated questionnaires will supersede the methods outlined herein.

12.9 Additional sensitivity analyses

12.9.1 Sensitivity of results to potential imbalance of baseline characteristics

The randomisation of a relatively large sample of participants into groups (anticipating approximately 90 participants per treatment group in this study) ensures that the average values of individual's characteristics at baseline are similar across the groups that is, they are balanced. Imbalances, however, are likely to occur by chance and when they do, they affect the inference about the treatment effect.

In this study, the following baseline variables will be examined for chance imbalances: 1) education level, 2) age, 3) ethnicity, 4) paranoia severity, and 5) bias score (SRT). To assess the effect of potential imbalance all five variables will be controlled for and included as additional covariates in the models. Any changes in treatment differences will be reported.

12.9.2 Sensitivity of results to the increase in precision

The five variables mentioned in the previous paragraph were chosen because they were thought to be important in predicting the outcome and increase the precision of the treatment effects. To assess the sensitivity of the standard errors in the presence of these variables, we will re-run and report the findings after adding 1) education level, 2) age, 3) ethnicity, and 4) SRT to the primary model, and 1) education level, 2) age, 3) ethnicity, 4) SRT, and 5) paranoia severity to the secondary models.

12.9.3 Sensitivity analysis due to Protocol violations

A protocol violation or non-compliance is any deviation from the protocol. The statistical analyses are concerned in particular with two protocol violations:

- Violations with regards to the occurrence of a session and the extent of the adherence to the study, as described in Section Strategies to Improve Adherence to Interventions.
- Violations due to unblinding.

We will assess violations to the protocol by assessing the violations (i.e. Unblinding or data was collected outside the collection time frame) in further sensitivity analysis by supplementing the model of the primary analysis, described above, with an additional binary indicator variable coded 1 for existence of any Protocol violation (i.e. unblinding at follow-up time), and 0 otherwise. The analysis of this model will give intervention effect estimates adjusted for potential effects of Protocol violation. We will report the changes in the predicted outcome differences alongside the main analysis. Any other protocol violation will be handled in the same way.

12.10 Exploratory dose-response analyses

The study design includes 12 weekly post-session interim assessments. This permits us to develop a dose-response model to examine the effect of treatment on paranoia growth by allowing a retrospective analysis of the primary outcome over time as the therapy progresses.

The dose-response model will be non-linear in the parameters and will take the form $y_{ij} = f(x_{ij};\ \beta, u_j) + \varepsilon_{ij}$, where y_{ij} is our primary response, paranoia scale, for time i and for participant j, and ε_{ij} are the errors. The function $f(x_{ij};\ \beta, u_j)$, which holds a covariate matrix x_{ij}, a vector of fixed effects β, and a vector of random effects u_j, will be a three-parameter logistic growth function (Ritz & Streibig, 2005). The choice of this popular growth function, in which participants' profiles of growth resemble an S shape, is based on the data we obtained

during the feasibility study. The non-linear model with the three-parameter logistic function takes the form:

$$paranoia_{ij} = \varphi_{1j}/(1 + \exp(-(dose_{ij} - \varphi_{2j})/\varphi_{3j})) + \varepsilon_{ij}$$

where dose will be defined as the number of once-a-week sessions, with a maximum of 12 sessions. The parameter φ_{1j} denotes the average paranoia score as dose, or the number of sessions go to infinity. The parameter φ_{2j} denotes the session at which half the average paranoia score is reached, in the sense that if $dose_{ij} = \varphi_{2j}$ then $E(paranoia_{ij}) = 0.5\varphi_{1j}$. Finally, φ_{3j} is a scale parameter which represents the number of weeks it takes for paranoia to move from 50% to about 73% of its growth.

To explore whether and how the treatment arms affect the growth of paranoia, we will equate each of the three parameters, φ_{1j}, φ_{2j}, and φ_{3j} to treatment. In particular, each parameter will be equated to treatment arm, controlling covariates (baseline outcome, study site and sex at birth), and a random intercept parameter, to allow for the growth to be participant specific. We, therefore, write:

$$\begin{aligned}\varphi_{1j} &= constant_1 + b_1 treatment_j + b_2 base.paranoia_j + b_3 gender_j + b_4 study.site_j + u_{1j}\varphi_{2j} \\ &= constant_2 + b_5 treatment_j + b_6 base.paranoia_j + b_7 gender_j \\ &\quad + b_8 study.site_j + u_{2j} \\ \varphi_{3j} &= constant_3 + b_9 treatment_j + b_{10} base.paranoia_j + b_{11} gender_j + b_{12} study.site_j + u_{3j}\end{aligned}$$

The treatment effect b_1 will enable us to assess whether there is a (adjusted) treatment difference in the average paranoia score that is obtained with higher doses. b_5 will denote the difference in weeks between the treatments to reach half of the expected maximum paranoia score, and b_9 the difference in weeks the treatments take for the paranoia score to go from 50% to 73% of its maximum growth. For example, if $b_1 = 4$ this means that the average paranoia score at high doses is 4 points greater for treatment 1 than for treatment 0, and if $b_5 = -2$ this means that treatment 0 reaches 50% of the paranoia score two weeks faster than treatment 1.

The dose-response model will be fitted using R's interface to JAGS package *rjags* (Plummer, Stukalov & Denwood, 2021), a program for the analysis of Bayesian hierarchical models using Markov Chain Monte Carlo simulation. The response will be assumed to follow a Normal distribution with mean $f(x_{ij};\ \beta, u_j)$ and inverse-variance in terms of a standard deviation that will be given an uninformative prior Gamma(0.01, 0.02) distribution. The rest of the model parameters will also be assigned non-informative priors. In particular, the constants $constant_1, constant_2$ and $constant_3$ that denote, respectively, 1) the expected maximum average paranoia score, 2) the week at which participants reach half of their expected maximum average paranoia score and 3) the expected weeks needed for the average paranoia score per participant to go from 50% to 73% of the limiting growth, will be assumed to follow a Uniform distribution on the range 1 to 100. The covariate effects b_1,…,b_{12} will be assigned Uniform distributions on the range -100 to 100. The covariance matrix of the random intercepts will be modelled directly using the inverse Wishart distribution.

Initially, we will allow 3 to 4 MCMC chains each one being iterated 80,000 times with a burn-in period of 100,000 and default initial values for all parameters. Diagnostic tests of the MCMC chains will include the graphical examination and the efficiency of the sampler. We will consider a sampler as efficient if the number of accepted proposals of model parameters relative to the total number of proposals is between 15% and 50% ensuring low autocorrelation and a relatively large effective sample size for all model parameters. In case of non-convergence or low effective sample size, we will consider 1) extending the chains and the burn-in period, 2) change default values of the parameters, 3) simplify or re-parameterise the model.

As a sensitivity analysis, several other non-linear functions will be explored using the R package *drc* (Ritz & Streibig, 2016) such as Gompertz, Weibull and exponential functions. The *drc* package does not allow for the use of repeated measurements and hence, we will select 10-20% of the study participants randomly and fit a non-linear model to each one of them separately. The models will then be compared through graphical means Graphs will predominately be trellis plots displaying paranoia score versus dose and the model fit for each participant individually. If we find evidence of a better fit with a different non-linear function, we will change the mean of the Normal likelihood of the main Bayesian model to

accommodate the new function and compare models using the deviance information criterion.

12.11 Planned subgroup analyses

There are no powered subgroup analyses planned.

12.12 Interim analysis

No interim analysis is planned in this study. If recruitment, adherence or other milestones are not met the DMC and TSC may request interim analyses to help inform their recommendations to the funder (MRC) about whether the study should stop or continue

13 Software

Data will be exported from the CTU MACRO database into Stata file format (.dta). All data processing and statistical analyses for the main trial paper will be performed using Stata versions 16 or higher.

14 Access to Protocol, participant level-data and statistical code

The trial protocol will be published in an academic journal and linked from the ISRCTN registry. If not published as an appendix of the trial protocol, the statistical analysis plan will be deposited at the ISRCTN registry once approved by the Trial Steering Committee. This is a clinical trial registry recognised by WHO and ICMJE that accepts planned, ongoing or completed studies of any design. It provides content validation and curation and the unique identification number necessary for publication. All study records in the database are freely accessible and searchable. The quantitative data generated by this study, as well as the statistical code, will be deposited with UK Data Archive via the ReShare process. See the Data Management Plan (Appendix 2 of the trial protocol) for more details on data sharing and

access. The statistical code will be deposited in GitHub (www.github.com), a Git repository hosting service.

15 References

Appelbaum PS, Robbins PC, Roth LH. Dimensional approach to delusions: comparison across types and diagnoses. Am J Psychiatry. 1999; 156(12):1938-43.

Bentall RP, Rowse G, Shryane N, Kinderman P, Howard R, Blackwood N, Moore R, Corcoran R. The cognitive and affective structure of paranoid delusions: a transdiagnostic investigation of patients with schizophrenia spectrum disorders and depression. Arch Gen Psychiatry. 2009;66:236-47.

Birchwood M, Trower P. The future of cognitive–behavioural therapy for psychosis: not a quasi-neuroleptic. Br J Psychiatry. 2006;188:107-8.

Carpenter, J., Kenward, M. Missing data in randomised controlled trials - a practical guide, 2007. https://researchonline.lshtm.ac.uk/id/eprint/4018500/

Carpenter, J. R., Roger, J. H. & Kenward, M. G. (2013). Analysis of Longitudinal Trials with Protocol Deviation: A Framework for Relevant, Accessible Assumptions, and Inference via Multiple Imputation. Journal of Biopharmaceutical Statistics 23, 1352-1371.

Carpenter, J. and Smuk, M. (2021) Missing data: A statistical framework for practice. Biometrical Journal 63, 915-947.

Castle DJ, Phelan M, Wessely S, Murray RM. Which patients with non-affective functional psychosis are not admitted at first psychiatric contact? Br J Psychiatry. 1994;165:101-6.

Fenigstein, A., & Vanable, P. A. (1992). Paranoia and self-consciousness. *Journal of Personality and Social Psychology, 62*(1), 129–138. https://doi.org/10.1037/0022-3514.62.1.129

Freeman D, Garety PA, Kuipers E, Fowler D, Bebbington PE. A cognitive model of persecutory delusions. Br J Clin Psychol. 2002;41:331-47.

Freeman D, Garety PA, Bebbington PE, Smith B, Rollinson R, Fowler D, Kuipers E, Ray K, Dunn G. Psychological investigation of the structure of paranoia in a non-clinical population. Br J Psychiatry. 2005;186:427-435.

Freeman D, McManus S, Brugha T, Meltzer H, Jenkins R, Bebbington P. Concomitants of paranoia in the general population. Psychol Med. 2011;41:923-36.

Freeman, D., et al., The Dunn Worry Questionnaire and the Paranoia Worries Questionnaire: new assessments of worry. Psychological Medicine, 2020. 50(5): p. 771-780.

Freeman, D., Loe, B., Kingdon, D., Startup, H., Molodynski, A., Rosebrock, L., Brown, P., Sheaves, B., Waite, F., Bird, J. (2021). The revised Green et al., Paranoid Thoughts Scale (R-GPTS): Psychometric properties, severity ranges, and clinical cut-offs. *Psychological Medicine, 51*(2), 244-253.

Garety PA, Kuipers E, Fowler D, Freeman D, Bebbington P. A cognitive model of the positive symptoms of psychosis. Psychol Med. 2001;31:189-95.

Goodwin FK, Jamison KR. Manic-depressive illness: bipolar disorders and recurrent depression. New York: Oxford University Press; 2007.

Hamner MB, Frueh BC, Ulmer HG, Arana GW. Psychotic features and illness severity in combat veterans with chronic posttraumatic stress disorder. Biol Psychiatry. 1999;45:846-52.

Harville, D.A. Maximum likelihood approaches to variance component estimation and to related problems, *Journal of the American Statistical Association*. 1977;72: 358

International Council for Harmonisation (2019). ICH E9 (R1) Addendum on estimands and sensitivity analysis in clinical trials to the guideline om statistical principles for clinical trials E9(R) (https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf). 2019.

Johnson J, Horwath E, Weissman MM. The validity of major depression with psychotic features based on a community study. Arch Gen Psychiatry. 1991;48:1075-81.

Kahan, B.C., Morris, T.P., White, I.R. et al. Estimands in published protocols of randomised trials: urgent improvement needed. Trials 22, 686 (2021). https://doi.org/10.1186/s13063-021-05644-4
Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophr Bull. 1987;13(2):261-76.

Keetharuth AD, Brazier J, Connell J, Bjorner JB, Carlton J, Taylor Buck E, Ricketts T, McKendrick K, Browne J, Croudace T, Barkham M. Recovering Quality of Life (ReQoL): a new generic self-reported outcome measure for use with people experiencing mental health difficulties. Br J Psychiatry. 2018 Jan;212(1):42-49.

Mallinckrodt, C.H., et al. Accounting for dropout bias using mixed-effects models, *Journal of Biopharmaceutical Statistics*, 2001;11: 9-21

Mathews, A. and B. Mackintosh, Induced emotional interpretation bias and anxiety. Journal of Abnormal Psychology, 2000. 109(4): p. 602-615.

Moritz S, Woodward TS. Metacognitive training in schizophrenia: from basic research to knowledge translation and intervention. Curr Opin Psychiatry. 2007;20:619-25.

Neil, S.T., et al., The questionnaire about the process of recovery (QPR): A measurement tool developed in collaboration with service users. Psychosis-Psychological Social and Integrative Approaches, 2009. 1(2): p. 145-155.

Plummer, M., Stukalov, A., Denwood, M. Bayesian graphical models using MCMC, R package version 4-11, 2021. https://cran.r-project.org/web/packages/rjags/rjags.pdf

Perälä J, Suvisaari J, Saarni SI, Kuoppasalmi K, Isometsä E, Pirkola S, Partonen T, Tuulio-Henriksson A, Hintikka J, Kieseppä T. Lifetime prevalence of psychotic and bipolar I disorders in a general population. Arch Gen Psychiatry. 2007;64:19-28.

Ritz, C., Streibig, J.C. (2005) Bioassay analysis using R, *Journal of Statistical Software*, 12: 5

Ritz, C., Streibig, J.C. (2016) drc: Analysis of dose-response curves, R package version 3.0-1, https://cran.r-project.org/web/packages/drc/drc.pdf

Rubin, D.B. Multiple Imputation for Nonresponse in Surveys, John Wiley and Sons, 1987, USA

Rude, S.S., et al., Negative cognitive biases predict subsequent depression. Cognitive Therapy and Research, 2003. 27(4): p. 415-429.

Savulich G, Freeman D, Shergill S, Yiend J. Interpretation biases in paranoia. Behav ther. 2015;46:110-24.

Senn, S. J. (2017). Contribution to the discussion of “A critical evaluation of the current p-value controversy.” Biometrical Journal. DOI: https://doi.org/10.1002/bimj.201700032

Tennant, R., et al., The Warwick-Edinburgh mental well-being scale (WEMWBS): development and UK validation. Health and Quality of Life Outcomes, 2007. 5.

Van Os J, Verdoux H, Maurice-Tison S, Gay B, Liraud F, Salamon R, Bourgeois M. Self-reported psychosis-like symptoms and the continuum of psychosis. Soc Psychiatry Psychiatr Epidemiol. 1999;34:459-63.

Waller H, Freeman D, Jolley S, Dunn G, Garety P. Targeting reasoning biases in delusions: a pilot study of the Maudsley Review Training Programme for individuals with persistent, high conviction delusions. J Behav Ther Exp Psychiatry. 2011;42:414-21.

Wessely S, Buchanan A, Reed A, Cutting J, Everitt B, Garety P, Taylor P. Acting on delusions. I: Prevalence. Br J Psychiatry. 1993;163:69-76.

White, I. R., Thompson, S. G. Adjusting for partially missing baseline measurements in randomized trials, *Statistics in Medicine*: 2005;993-1007.

White, Ian R. (2011) Strategies for handling missing data in randomised trials. Trials 12, Suppl 1 A59. 2011, doi:10.1186/1745-6215-12-S1-A59

Yiend J, Trotta A, Meek C, Dzafic I, Baldus N, Crane B, Kabir T, Stahl D, Heslin M, Shergill S, McGuire P, Peters E. Cognitive bias modification for paranoia (CBM-pa): Study protocol for a randomised controlled trial. Trial. 2017;18:298.

Zigmond A.S., Snaith R.P. The hospital anxiety and depression scale. Acta Psychiat Scand. 1983;67:361-70.