MUSE-FEP study
Statistical Analysis Plan

Appendices

James Liley, Ehsan Kharatikoopaei, Emmanuel Ogundimu

Table of Contents

1 Appendix 1 (detailed MUSE-FEP scoring guidance)

1.1 PSYRATS scoring

The Psychotic Symptom Rating Scales (PSYRATS (Haddock et al., 1999)) is a clinician administered semi-structured interview of hallucinations (such as amount/intensity of distress). It consists of 11 items, scored from 0-4. PSYRATS will be used to assess the multidimensional aspects of hallucinations (such as distress, preoccupation, and conviction). There are 11 items for hallucinations, with 5 of the items used to identify voice related distress. There are six items for delusions (Woodward et al., 2014). PSYRATS is well suited to assess outcome in psychological therapies and has been used in major treatment trials. The total score and the sums of scores across subscales for the PSYRATS will be reported, with a particular focus on Voice related distress and Total score for PSYRATS AH (11 items)

PSYRATS hallucination subscales are defined as follows

- Voice related distress subscale H-DIS (items 6,7,8,9 11)
- Frequency H-FREQ (items 1,2, 10)
- Attribution H-Attr Cognitive (items 3 and 5)
- Loudness H-Loud (item 4)

In addition, the total score for PSYRATS Delusions (6 items) will be reported.

1.2 Hamilton score

The Hamilton Program for Schizophrenia Voices Questionnaire (HPSVQ, (Van Lieshout & Goldberg, n.d.)) is a patient-reported questionnaire on auditory hallucinations. It has nine items rated on a 5-point scale, which are summed to a total score. The 'negative impact' subscale comprises the sums of items 2,5,6,7 and measures the level of distress and impact that voices have on the person. The overall total score and the total score across the negative impact subscale will be reported.

1.3 DASS score

The short Depression, Anxiety and Stress Scales (DASS; (Lovibond & Lovibond, 1995)) is a 21 item self -report questionnaire designed to assess symptoms of anxiety, depression and stress. It comprises 21 items, scored from 0 to 3, with a higher score indicating worse symptoms. It has a range of subscales. The total score and totals across subscales will be reported.

The subscales comprise the following items:

- Stress 1,6,8,11,12,14,18

- Anxiety 2,4,7,9,15,19,20
- Depression 3,5,10,13,16,17,21

1.4 QPR score

The Questionnaire for the Process of Recovery QPR (Neil et al., 2009) is a user-defined measure, assessing subjective recovery in intrapersonal and interpersonal functioning. The QPR has 15 items each scored on a 4-point scale (0= disagree strongly, 1=disagree, 2=neither agree nor disagree, 3=agree, 4=agree strongly). Higher scores are indicative of recovery. We will report the total QPR score.

1.5 CHOICEs score

The CHoice of Outcome In Cbt for psychosEs (CHOICE; (Greenwood et al., 2010)), is a 21 item service-user developed questionnaire to evaluate outcomes for people with psychosis and assess therapy-related goals. Study participants completed the long version, but we used a short version (derived from the long form) which comprises 11 items (1, 2, 3, 7, 9, 12, 17, 18, 20, 21, 22). The short form also includes a free-text personally defined goal, which we did not use for statistical analysis. This short form is reversed scored from the long form where higher scores indicate worse wellbeing/functioning. In the short form higher scores indicate better outcomes.

1.6 SF-36 score

The SF-36 score (Ware Jr & Sherbourne, 1992) measures general activity and wellbeing in the past week. It comprises 36 items , with a higher score corresponding to better wellbeing. The range of possible responses varies across items.

Items are organised into subscales as indicated below.

- Physical functioning: 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
- Role limitations due to physical health: 13, 14, 15, 16
- Role limitations due to emotional problems: 17, 18, 19
- Energy/fatigue: 23, 27, 29, 31
- Emotional well-being: 24, 25, 26, 28, 30
- Social functioning: 20, 32
- Pain: 21, 22
- General health: 1, 33, 34, 35, 36

The score for each subscale is computed as a weighted score of responses for items in that subscale. Specifications of weights are as follows. The notation (*a,b,c*) - (*d,e,f*) means *d* is substituted for *a*, *e* for *b*, and *f* for *c*. After substitutions, the substituted values are summed within the subscales above to give a score for each subscale.

- Questions 1, 2, 20, 22, 34, 36: (1,2,3,4,5) - (100,75,50,250,0)
- Questions 3, 4, 5, 6, 7, 8, 9, 10, 11, 12: (1,2,3) - (0,50,100)
- Questions 13, 14, 15, 16, 17, 18, 19: (1,2) - (0,100)
- Questions 21, 23, 26, 27, 30: (1,2,3,4,5,6) - (100,80,60,40,20,0)
- Questions 24, 25, 28, 29, 31: (1,2,3,4,5,6) - (0,20,40,60,80,100)
- Questions 32, 33, 35: (1,2,3,4,5) - (0,25,50,75,100)

The scoring protocol is given in detail at:

https://www.rand.org/health-care/surveys_tools/mos/36-item-short-form/scoring.html

The Physical component summary (PCS) and mental component summary (MCS) are computed as weighted sums of subscales. The weights and intercepts are given in Table 1

SF-36 scale	PCS	MCS
Physical Functioning (PF)	0.18520282	-0.100454686
Role Physical (RP)	0.10391070	-0.036479256
Bodily Pain (BP)	0.13478621	-0.041305177
General Health (GH)	0.12372060	-0.007788934
Vitality (VT)	0.01378572	0.112767868
Social Functioning (SF)	-0.00336515	0.120108579
Role Emotional (RE)	-0.05815212	0.131428154
Mental Health (MH)	-0.12252462	0.269716282
Intercept (added)	20.13602	17.33727

Table 1: PCS and MCS transform

1.7 ICECAP score

The ICECAP-A score (Investigating Choice Experiments Capability Measure for Adults, (Flynn et al., 2015)) consists of five domains (Secure, Support, Independence, Achievement, Enjoyment) scored on a 1-4 scale, with 4 being better. We will report total score.

1.8 EQ-5D score

The EQ-5D score (EUROQOL-5D, (Herdman et al., 2011)) comprises five domains (mobility, self care, usual activities, pain/discomfort, anxiety and depression) and a rating of current perception of health. We will consider each domain separately. Higher scores indicate worse perceived health. We will not use EQ-5D directly as an outcome in the study. Further information on the EQ-5D score is available at:

https://euroqol.org/publications/user-guides/

1.9 Satisfaction with therapist

We will use the therapeutic alliance & therapy acceptability (Oei & Green, 2008) which is a short scale assessing overall acceptability of the therapeutic interaction. The score comprises subscales assessing satisfaction with therapy and satisfaction with therapist, and a further single score assessing global improvement. The subscale indicating satisfaction with therapy is all even-numbered questions, and the subscale indcating satisfaction with the therapist is all odd-numbered questions excluding question 13. We will separately report the two subscales and question 13 which is a patient's view of impact of therapy (lower scores indicate better outcome on this version).

1.10 Working alliance inventory

The working alliance inventory (Horvath & Greenberg, 1989) assesses alliance between therapist and patient. We will use a short form of the questionnaire (Hatcher & Gillaspy, 2006). Results from this measure will be used for qualitative analysis only.

1.11 Service use measure

The medical records of all participants will be checked at 8 weeks and 12 weeks to determine a) whether the person has had adverse or serious adverse events, b) has had any period of hospitalisation, c) any changes in medication, d) whether they have attended MUSE sessions (in TAU or MUSE arms) and d) whether it is possible to determine service use from the medical records. This latter part is an assessment of whether use of GP, education, work, and reliance on family or carers is noted in the records. It is descriptive and used to help determine if we can access this information from medical records. The measure overall will determine number of adverse and serious adverse events in each group, and number of sessions of MUSE attended.

2 Appendix 2: statistical details

2.1 Linear mixed model for assessing effect sizes

We will index individuals by $i \in 1,2,\ldots,N$. We will index time by $t \in 1,2,3$ where 1 is baseline, 2 is end-of-treatment (approximately 8 weeks) and 3 is follow-up (approximately 12-16 weeks).

Our model is as follows:

$$Y_{it} = \beta_0 + \beta_{1a}1_{t=1} + (\beta_{2a} + \beta_2 u_i)1_{t=2} + (\beta_{3a} + \beta_3 u_i)1_{t=3} + X_i\beta + \gamma_i + \varepsilon_{it}$$

where

- 1_C is 1 if condition C is satisfied and 0 otherwise
- Y_{it} is the outcome measurement for individual i at time t
- u_i is a treatment indicator for the the ith individual: 1 if treated, 0 if not.
- β_0 is an intercept common to all individuals,
- $\beta_{1a}, \beta_{2a}, \beta_{3a}$ are fixed intercepts for each time point,
- β_2 is the effect of treatment on the outcome at $t = 2$ (end-of-treatment)
- β_3 is the effect of treatment on the outcome at $t = 3$ (follow-up)
- β is a vector of parameters for fixed effects
- X_i is a vector of indicator values for fixed effects
- γ_i is a random intercept to for the ith individual. The values γ_i are independently and identically distributed as $N(0, \sigma_r^2)$.
- ε_{it} is a random error term. The values ε_{it} are independently and identically distributed as $N(0, \sigma_e^2)$.

We will make maximum-likelihood estimates and compute associated asymptotic standard errors and confidence intervals for parameters β_0 , β_{1a}, β_{2a}, β_{3a} , β_2, β_3 , β, σ_r and σ_e. Our main parameter of interest is β_2.

2.2 Considerations for estimated minimal sample sizes for a definitive study

A partial aim of this study is to assist in the design of a future definitive trial to evaluate the effect of MUSE-FEP on some outcome.

Two essential aspects of such planning are the choice of primary outcome and an estimate of the minimum sample size necessary to detect a given effect on this primary outcome with sufficient power.

The degree of effect which the definitive trial will be powered to detect is generally a question of a minimal clinically significant difference in the outcome in question. However, in determining such a degree of effect for computing a minimum sample size, an estimate of the true effect of MUSE-FEP on the outcome is useful.

For this reason, we wish to obtain an essentially unbiased estimate Z of the true effect $Z_{true} = E\{\beta_2/SE(\beta_2)\}$ for the outcome we choose to use as the primary outcome in the definitive trial. We may also consider the minimum sample size of a definitive trial to detect an effect the size of the (estimated) Z. We will presume that the linear-mixed-model approach in the previous section gives an essentially unbiased estimate of Z_{true}.

A difficulty immediately encountered is that if we were to select a primary outcome on the basis of the largest estimate Z in this trial, we would induce an upward bias in our estimate of Z_{true} for that effect due to 'winner's curse', or 'regression to the mean' (Barnett et al., 2005; Galton, 1886). An implication of this is that an estimate of minimum sample size to detect an effect the size of estimate Z for this outcome will (generally) be biased downwards compared to the minimum sample size to detect an effect the size of Z_{true}.

We manage this difficulty by nominating a pseudo-primary outcome prior to seeing the data. Under an assumption that the true value of Z_{true} is equal for all potential primary outcomes, the estimate of Z for the pseudo-primary outcome will be an unbiased estimate of Z_{true}. The minimum sample size corresponding to Z will be an essentially unbiased estimate of the minimum sample sample size necessary to detect Z_{true}, discounting biases due to non-linearity of the relation between effect size and minimum sample size.

2.3 Note on inference and description

In this study, we are avoiding inference on effect sizes, so we are not identifying nor rejecting null hypotheses. Our adherence to this principle will not be results - dependent: we will not claim that the study shows that MUSE has an effect on outcome, whatever the result.

Our agnosticism to results in this principle is important in avoidance of publication bias (Sackett, 1979).

3 Appendix 3. CONSORT guidelines checklist

In this appendix, we report concordance with the CONSORT guidelines for publication of a pilot trial (Eldridge et al., 2016). Not all elements are relevant to this SAP, and we have marked irrelevant elements as 'NA'.

Section/topic	Num.	Extension for pilot trials	Section
Title/abstract	1a	Identification as a pilot or feasibility randomised trial in the title	NA
	1b	Structured summary of pilot trial design, methods, results, and conclusions (for specific guidance see CONSORT abstract extension for pilot trials)	4.1
Introduction	2a	Scientific background and explanation of rationale for future definitive trial, and reasons for randomised pilot trial	4.1, 4.2
	2b	Specific objectives or research questions for pilot trial	5
Trial design	3a	Description of pilot trial design (such as parallel, factorial) including allocation ratio	6
	3b	Important changes to methods after pilot trial commencement (such as eligibility criteria), with reasons	11.6, 11.7, 11.8
Participants	4c	How participants were identified and consented	Protocol
Outcomes	6a	Completely defined prespecified assessments or measurements to address each pilot trial objective specified in 2b, including how and when they were assessed	8.2, appendix 1
	6b	Any changes to pilot trial assessments or measurements after the pilot trial commenced, with reasons	11.8
	6c	If applicable, prespecified criteria used to judge whether, or how, to proceed with future definitive trial	11.7

Sample size	7a	Rationale for numbers in the pilot trial	9
Sequence generation	8b	Type of randomisation(s); details of any restriction (such as blocking and block size)	7
Analytical methods	12a	Methods used to address each pilot trial objective whether qualitative or quantitative	10.2,12
Participant flow	13a	For each group, the numbers of participants who were approached and/or assessed for eligibility, randomly assigned, received intended treatment, and were assessed for each objective	6, Figure 1
Recruitment	14b	Why the pilot trial ended or was stopped	9, 10.8
Numbers analysed	16	For each objective, number of participants (denominator) included in each analysis. If relevant, these numbers should be by randomised group	10.2, 12
Outcomes and estimation	17a	For each objective, results including expressions of uncertainty (such as 95% confidence interval) for any estimates. If relevant, these results should be by randomised group	10.2, 12, Appendix 2
Ancillary analyses	18	Results of any other analyses performed that could be used to inform the future definitive trial	10.2, 12, Appendix 2
Harms	19a	If relevant, other important unintended consequences	10.2, 12, Appendix 2
Limitations	20	Pilot trial limitations, addressing sources of potential bias and remaining uncertainty about feasibility	Appendix 2
Generalisability	21	Generalisability (applicability) of pilot trial methods and findings to future definitive trial and other studies	Appendix 2
Interpretation	22	Interpretation consistent with pilot trial objectives and findings, balancing potential benefits and harms, and considering other relevant evidence	NA
	22a	Implications for progression from pilot to future definitive trial, including any proposed amendments	11.7

Registration	23	Registration number for pilot trial and name of trial registry	Frontmatter
Protocol	24	Where the pilot trial protocol can be accessed, if available	Frontmatter
Funding	26	Ethical approval or approval by research review committee, confirmed with reference number	Frontmatter

4 References

Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: What it is and how to deal with it. *International Journal of Epidemiology*, *34*(1), 215–220. https://doi.org/10.1093/ije/dyh299

Eldridge, S. M., Chan, C. L., Campbell, M. J., Bond, C. M., Hopewell, S., Thabane, L., & Lancaster, G. A. (2016). CONSORT 2010 statement: Extension to randomised pilot and feasibility trials. *Bmj*, *355*.

Flynn, T. N., Huynh, E., Peters, T. J., Al-Janabi, H., Clemens, S., Moody, A., & Coast, J. (2015). Scoring the ICECAP-A capability instrument. Estimation of a UK general population tariff. *Health Economics*, *24*(3), 258–269.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, *15*, 246–263.

Greenwood, K. E., Sweeney, A., Williams, S., Garety, P., Kuipers, E., Scott, J., & Peters, E. (2010). CHoice of Outcome In Cbt for psychosEs (CHOICE): The development of a new service user–led outcome measure of CBT for psychosis. *Schizophrenia Bulletin*, *36*(1), 126–135.

Haddock, G., McCarron, J., Tarrier, N., & Faragher, E. B. (1999). Scales to measure dimensions of hallucinations and delusions: The psychotic symptom rating scales (PSYRATS). *Psychological Medicine*, *29*(4), 879–889.

Hatcher, R. L., & Gillaspy, J. A. (2006). Development and validation of a revised short version of the Working Alliance Inventory. *Psychotherapy Research*, *16*(1), 12–25.

Herdman, M., Gudex, C., Lloyd, A., Janssen, M. F., Kind, P., Parkin, D., Bonsel, G., & Badia, X. (2011). Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Quality of Life Research*, *20*(10), 1727–1736.

Horvath, A. O., & Greenberg, L. S. (1989). Development and validation of the Working Alliance Inventory. *Journal of Counseling Psychology*, *36*(2), 223.

Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. *Behaviour Research and Therapy*, *33*(3), 335–343.

Neil, S. T., Kilbride, M., Pitt, L., Nothard, S., Welford, M., Sellwood, W., & Morrison, A. P. (2009). The questionnaire about the process of recovery (QPR): A measurement tool developed in collaboration with service users. *Psychosis*, *1*(2), 145–155.

Oei, T. P., & Green, A. L. (2008). The Satisfaction With Therapy and Therapist Scale–Revised (STTS-R) for group psychotherapy: Psychometric properties and confirmatory factor analysis. *Professional Psychology: Research and Practice*, *39*(4), 435.

Sackett, D. L. (1979). Bias in analytic research. In *The case-control study consensus and controversy* (pp. 51–63). Elsevier.

Van Lieshout, R. J., & Goldberg, J. O. (n.d.). Hamilton Program for Schizophrenia Voices Questionnaire. *Canadian Journal of Behavioural Science/Revue Canadienne Des Sciences Du Comportement*.

Ware Jr, J. E., & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36): I. Conceptual framework and item selection. *Medical Care*, 473–483.

Woodward, T. S., Jung, K., Hwang, H., Yin, J., Taylor, L., Menon, M., Peters, E., Kuipers, E., Waters, F., & Lecomte, T. (2014). Symptom dimensions of the psychotic symptom rating scales in psychosis: A multisite study. *Schizophrenia Bulletin*, *40*(Suppl_4), S265–S274.