PARADES PSYCHOEDUCATION TRIAL

Statistical Analysis Plan Version 1.1

Fiona Holland, Chris Roberts

August 18th 2014

This document considers the statistical analyses for the Parades Psychoeducation Trial which compares a Psychoeducation Group (PE) with a Bipolar Peer Support Group (PS). The purpose of the SAP is to document the confirmatory statistical analyses of the trial thereby controlling for statistical analyses bias. The statistical analyses follow the principles of ICH E9.

1. Background

There is evidence that group psychoeducation is more effective than group support in a specialist mood disorder centre in Spain[1,2] (with benefits sustained for up to 5 years) and treatment as usual in Austrialia[3,4]. Group psychoeducation has not been formally evaluated in the UK, and there are no randomised controlled trials examining the effectiveness of jointly delivered expert patient and professional group psycheducation.

The main aim of this trial is to determine whether psychoeducation is more clinically and cost effective than unstructured peer support in the management of bipolar disorder.

2. Study Design

The Parades psychoeducation trial is a single blind two centre randomised controlled trial of 21 sessions of group psycheducation plus treatment as usual (TAU) versus 21 sessions of group peer support in adults with bipolar disorder plus TAU, for people not in a current episode but who have relapsed at least once in the previous two years. For both PE and PS, each session is attended by two health professionals (nurse, psychiatrist, psychologist or occupational therapist) and one of a small group of expert patients/service users (from a pool of expert patients trained for the purpose). In the PS group the health professionals and expert patient will be more passive being there to facilitate discussion and encourage participation and to correct any false information plus monitor risk in the group in the same way as in the PE group.

Sample size: Assuming a small clustering effect of 0.05 and a differential treatment effect of 0.22 (60% recurrence in the PS control group versus 38% in the PE group) and 80% power for detecting a difference at 0.05 significance level requires 82 participants per arm. Assuming a mean group size of 18 and an intra-class correlation coefficient of 0.05, a design effect of 1.85 gives a sample size of 152 in each arm. We assumed 15% loss to follow up giving a total sample size of 179, or a total of 358 randomised participants in total. Initially we assumed that this would

be achieved by running 10 groups of 18 participants in each arm (10 in the East Midlands and 10 in the North West region) over 3 years. However, due to lower numbers of people in each group this has been increased to 12 groups in the North West.

Allocation: Patients were individually randomised to PE or PS, with stratification by region (East Midlands/ North West) and minimisation in terms of number of previous episodes (<7, 8-19,20+).

3. Outcome Measures

3.1 Primary outcome measure

- Time from randomisation to the first week of recurrence of an episode of mania, hypomania, or a mixed episode for 1 week, or major depression for 2 consecutive weeks, satisfying DSM-IV criteria (both coded as a 5 or 6 on SCID LIFE, see below for interpretation)

The 16-weekly SCID Longitudinal Interval Follow-up Evaluation (LIFE) interviews (alternating between telephone and face to face interview) is used to generate weekly scores for mania and depression, both on a 1-6 scale of severity:

1=no symptoms,

2=minor symptoms

3=partial remission

4=does not meet DSM-IV criteria but major symptoms of the disorder

5=DSM-IV definite but no prominent psychotic symptoms and no extreme impairment in functioning

6=DSM-IV definite,severe and either prominent psychotic symptoms or extreme impairment in functioning

Those with a score of 5 or 6 on LIFE satisfy DSM-IV criteria. Note, if a participant consents but drops out from follow-up (or has a period of absence while on study) then the case notes (up to 96 weeks) are used to determine the time to recurrence, if applicable. Participants under the care of an NHS trust will have their electronic and paper notes reviewed whereas those seen only by their GP will have the GP notes reviewed.

Note, if an individual does not have a recurrence then they will be censored at 96 weeks. However, if the notes screening indicates that there has been no contact with the participant after a certain date then this will be their censor date, which will include date of death if applicable.

3.2 .Secondary outcome measures

- Time from randomisation to the first occurrence of a manic-type episode (mania, hypomania or mixed affective episode), satisfying DSM-IV criteria
- Time from randomisation to the first occurrence of a depressive episode, satisfying DSM-IV criteria

The following outcomes are all assumed to be continuous variables:

- Assessment of 16 weekly average mania and depression scores separately, using the SCID LIFE (see Section 7.2)
- Assessment of function using the Social Adjustment Scale (SAS) :overall score, and four domains: performance, interpersonal behaviour, friction and dependency
- Assessment of function using the *SOFAS*
- Hospital Anxiety and Depression Scale (HADS): anxiety and depression subscales
- The Short-Form 12 (SF-12) : physical (PCS) and mental component summaries (MCS)

A summary of the primary and secondary outcomes is shown in Table 1 below.

Table 1 Primary and secondary outcomes and number of inferential analyses

Measure	Number of outcomes	Description of components	Status	Number of inferential analyses
Time to first event	1	Mania or Depression	Primary	1
Time to first event	2	Mania and Depression separately	Secondary	2
SCID	2	Mania and Depression 16 weekly mean scores	Secondary	4= (main effect+ interaction)*2
SAS	5	overall score + 4 domains: performance, interpersonal behaviour, friction and dependency:	Secondary	10= (main effect+ interaction)*5
SOFAS	1	scale of 1-100	Secondary	2= (main effect+ interaction)
HADS	2	Anxiety and depression	Secondary	4= (main effect+ interaction)*2
SF-12	2	PCS, MCS	Secondary	4= (main effect+ interaction)*2
Total	**15**			**27**

3.3 Other Measures

- Hamilton Depression Rating Scale-GRID (HDRS)
- Beck-Raphaelson Mania Scale (MAS)
- The Social Rhythm Metric trait (SRM-t interview)

The following are self reported measures:

- The Hayward Stigma Questionnaire
- The Knowledge about Bipolar Disorder (KAB)
- The Hypomania Interpretations Questionnaire (HIQ): hypomanic, normalizing and experience
- The Social Rhythm Metric diary (SRM-d)
- The Early Warning Signs Checklist
- The Coping Strategies Checklist
- The Brief Illness Perception Questionnaire

In addition, the following information is available (see Section 8.1):

- Medication Adherence (MedAd)
- Treatment Fidelity which includes participant reports of what was covered in each session for PE and PS

3.4 Health Economic Outcomes

- Health status and related utility values using the Euroqol 5D and adapted CSRI
  [to be supplied by Linda Davies]

4. Data Sources

4.1 Pre randomisation data

Data are required for completion of the CONSORT diagram pre randomisation. These include:

- Numbers of potential participants assessed
- Numbers excluded after initial assessment by reason
- Numbers invited to baseline research interview
- Numbers excluded after baseline research interview by reason
- Numbers consenting and randomised by treatment arm

4.2 Baseline Data and Patients Characteristics for Randomised Patients

The information collected at baseline includes basic demographic and clinical data. The full list of baseline characteristics to be presented by treatment arm is given in Section 6.1. The SCID-LIFE and all the assessments listed in Table 2 below are performed at baseline. In addition the Group

Preference Scale and Expectations scale is completed only at baseline. Only baseline assessments prior to or on the day of randomisation will be used.

Table 2 Measures schedule up to 96 weeks and number of subscales where appropriate.

Baseline & Every 16 Weeks	Method	Subscales	Number of components
SCID – LIFE	interview	Mania and Depression	2
HDRS	interview		1
MAS	interview		1
MedAd	interview		N/A
Baseline & Every 32 Weeks			
SAS	interview	Overall, plus domains: performance, interpersonal behaviour, friction and dependency	5
SOFAS	interview		1
HADS	postal	Anxiety and depression	2
KAB	postal		1
EWS checklist	postal	Mania and Depression	2
Coping Strategies	postal	Mania and Depression	2
BIPQ	postal	Individual items	13
HIQ	postal	Hypomanic, normalizing and experience	3
Hayward Stigma	postal	Self-esteem and stigma	2
SF-12	postal	Physical and mental component summaries	2
{EuroQol	interview LD}		
{adapted CSRI	interview LD}		
{SRMetric trait	interview Lancs}		
{SRMetric diary	postal Lancs}		

{ }=undertaken by Linda Davies or at Spectrum, Lancaster University

4.3 Therapist Characteristics and Treatment Data

As suggested in the CONSORT guidance extension for trials of non-pharmacological intervention, information will be gathered regarding the characteristics of all health professionals and the expert patients for each intervention. The number of participants attending each group session, the therapists identifier and the type of workers e.g. nurse or psychiatrist conducting the session will be summarised. In addition, the number (percentage) of sessions attended (0,1,2,..,21) will be summarised.

4.4 Follow-up Assessments

The follow-up schedule is shown in Table 2. For the assessments done every 16 weeks telephone assessment are the basis of interviews at 16, 48 and 80 weeks. At 32, 64 and 96 weeks self-rated questionnaires are emailed or posted, as the participant requests.

5 Handling Missing Data and Slotting of Assessments

5.1 Item Non-response in Scale Measures

Item non-response for a scale or a subscale will be dealt with using a pro-rating strategy. Provided that at least 50% of the items are available the observed total (for the completed items) and the number of items completed will be used to calculate an adjusted total as follows:

Adjusted total = Observed total * Total number of items in scale/Number of items completed

Note, this is equivalent to replacing the missing item by the average of the participants available data for that dimension.

One exception will be the three HIQ scales where pro-rating will be performed provided that at least 80% (i.e.≥8 items) of the items are available.

Note, as individual items are considered for the BIPQ pro-rating will not be performed on this scale.

5.2 Missing baseline data

For inferential analyses, missing LIFE and other covariate data at baseline will be imputed using simple (deterministic) imputation[5] which is based on multiple regression. The following covariates listed in Section 7.1 will be used: sex, number of previous bipolar episodes (<7,8-19,20+, treated as a categorical variable) and wave (n=11). Baseline measurements taken after randomisation will not be considered for summary statistics and will rely on imputation of the baseline measure for inferential analysis. Substitution or imputation will not be used for post-baseline outcomes (see Section 7.2.3 for reasons).

5.3 Slotting of assessment measures

For the secondary outcome assessments (except for SCID LIFE) actual time from randomisation until assessment is likely to depart from the target assessment times. To provide summary statistics we need to assign each actual assessment to a target assessment week based on pre-defined intervals from time since randomisation. The assessment will be assigned to one of the following scheduled visit weeks based on the interval it falls into for the time (in weeks) from randomisation:

EVERY 16 WEEKS		EVERY 32 WEEKS	
Scheduled	Interval	Scheduled	Interval
0	<= rand date	0	<= rand date
16 weeks = [8 to <24]			
32 weeks = [24 to <40]		32 weeks = [16 to <48]	
48 weeks = [40 to<56]			
64 weeks = [56 to <72]		64 weeks = [48 to 80]	
80 weeks = [72 to <88]			
96 weeks = [88+]		96 weeks = [80+]	

Note these window intervals are arbitrary and are currently centered around the visit times.
If, for a given assessment, there is more than one measure in the band then the measurement nearest to the scheduled visit will be used for summary statistics.

6 Descriptive Analyses of randomised patients

6.1 Baseline Characteristics

Patients in the two treatment arms (PE and PS) will be described separately with respect to gender, age, ethnicity, marital status, living arrangements, number of children, education level and employment. In addition, current psychiatric treatment (including doses of drugs summing dose equivalents for polypharmacy), DSM-IV diagnosis, substance use, number of previous bipolar episodes (<7, 8-19,20+), presence of axis1 and axis 2 comorbid psychopathology, presence of borderline or antisocial personality disorder, suicide risk, and potential for violence will be presented. SCID LIFE mania and SCID LIFE depression scores will be summarised based on the previous four weeks,

Numbers (with percentages) for binary and categorical variables, and ordered categories plus means, standard deviations, medians plus minimums and maximums for continuous variables will be presented. Consistent with CONSORT guidance there will be no tests of statistical significance nor confidence intervals for differences between randomised groups on any baseline variable.

All baseline measurement scales will be summarised by treatment arm.

6.2 Follow-up

All measurement scales taken during follow-up will be summarised by time for each treatment arm. In addition, the number (percentage) of adverse events will be summarised by follow-up visit and treatment arm.

6.3 Loss to follow-up

The frequencies (with percentages) of patient losses to follow-up (i.e., drop-outs from the study) at 16, 32, 48, 64, 80 and 96 weeks after randomisation will be reported and compared between PE and PS arms. The reasons participants drop out of the study will be tabulated by treatment arm.

To capture the first post treatment and scheduled end of study period, treatment arm and baseline characteristics (see Section 6.1) of participants completing and not completing 32 and 96 weeks of

follow-up will be compared using logistic regression models. NB people should remain on the study once a recurrence of depression/mania occurs. This analysis will be used to develop an understanding of the missing data mechanism.

6.4 Observer Reliability between and within research sites

Intra and inter observer reliability will be considered using graphical methods and relevant summary statistics including intra-class correlation coefficients and kappa coefficients for the SCID LIFE measure.

7 Statistical analysis of outcomes

The analyses comparing PE with PS will be conducted applying the principle of intention to treat (ITT) subject to the availability of data. Statistical analysis of outcome will use a 5% two-sided significance level.

7.1 Primary Outcome – Comparison of recurrence between treatments

The primary outcome measure is time (in weeks) from randomisation to the first recurrence of an episode of major depression, hypomania, mania, or mixed episode, satisfying DSM-IV criteria. The number and percentage of patients satisfying criteria for a DSM-IV episode (at any time-point on the study), broken down by type (major depression; mania; hypomania or mixed) will be reported for each treatment arm and overall. The intervals (in weeks) from randomisation to recurrence (of any type) will be summarised by Kaplan-Meier curves with the estimated median times (if estimable) presented. A Cox proportional hazard model with robust variance estimators to account for nesting of participants within treatment groups (n=22) will be used to provide an estimate of the hazard ratio and 95% confidence interval. The following baseline covariates will be included in this model

- Sex
- Number of previous bipolar episodes prior to randomisation (<7, 8-19,20+ , as a categorical variable)
-
- Wave (n=11)
-
- Treatment allocation

Based on medical notes data, if appropriate those who have not relapsed by 96 weeks will be censored at 96 weeks and only relapses up to week 96 will be considered. In the absence of a

relapse, the censor date may be less than 96 weeks if no further contact information can be obtained from either the study data or medical notes.

In the presence of differential dropout (see section 6.3) a sensitivity analysis will be done on the 96 week relapse outcome data using logistic regression models.

7.1.1 Cox Model Diagnostics for the Primary Outcome

The Cox proportional hazards model assumes that the hazard ratio is constant over time. This assumption will be formally tested overall and additionally tested for each covariate. A natural log-log plot of survival against log time will be plotted for each group. Parallel lines implies the assumption of proportional hazards is satisfied.

Deviance residuals (a rescaling of Martingale residuals which are then symmetric about zero) whose behaviour is known approximately when the fitted model is satisfactory will be plotted. Measures of influence will also be used to identify participants who potentially have a large impact on the results.

If the model departs substantially from the assumption of proportional hazards or the residual or influence plots indicate marked lack of fit then an accelerated failure time (AFT) model will then be fitted and inference will be based on this alternative model. In the AFT model, the natural logarithm of the survival time, log t, is expressed as a linear function of the covariates where the effect of a covariate is to speed up or alternatively slow down the survival time.

7.2 Secondary Outcomes – Comparison of Mania and Depression Symptom scores

The SCID LIFE is rated every 16 weeks, retrospectively completing 16 weeks of scores. The main analysis of this data will estimate the treatment effect and the time with treatment interaction effect adjusted for baseline scores and other pre-specified baseline covariates. Follow-up data gathered in this way will tend to be more correlated within reporting period than between reporting period and may also be auto-correlated due to halo effects. One option would be to explicitly model this data structure, but this is unnecessary when estimating the treatment effect or the time with treatment interaction. A summary measure approach aggregating data for each reporting period will be more robust and equally efficient.

Each participants weekly data will be aggregated by calculating their average score over 16 weeks. This will provide an estimate of how each person felt over a period of 16 weeks corresponding to the period of recall. An alternative would be to consider, say, 4 week intervals. However if a shorter period such as this was chosen then there may be problem with autocorrelation. There would also

be variation in data quality as some 4 week periods would be based on data involving greater recall. An additional factor may be missing data and any overlap of periods. We therefore propose to carry out some analysis of the data after locking the database but before treatment allocation has been attached to decide the most appropriate period to aggregate, choose the method for dealing with missing data and any overlap of reporting periods.

In addition, the decision to aggregate over a 16 week interval was based on an analysis of the SCID LIFE data from an enhanced relapse prevention cluster randomised controlled trial[6]. On fitting models with a treatment by time interaction we found the estimated difference in slopes and treatment means were similar for outcome data based on 4 or 12 weekly SCID LIFE averages. Because auto-correlation is likely to be greater for the 4 weekly data compared to the 12 weekly averaged data the latter were preferred.

7.2.1 Summary of observed LIFE mania and depression scales

Summary statistics for the sixteen (or alternative) weekly averages will be tabulated by intervention and assessment time (including baseline). These averaged mania and depression SCID LIFE score profiles will also be plotted against time.

7.2.2 Linear Mixed Effect Modelling of Mania and Depression Symptom scores

For each person the average of their weekly scores over each 16- (or alternative) week interval post randomisation (1-16,17- 32, …, 81-96) will be calculated. Inferential analyses of the symptom scores (assumed to be normally distributed) will be based on linear mixed effects (LME, also know as random effects or random coefficient) models which include two parts: a) fixed main effects (or average response) and b) random effect terms accounting for the fact that measurements taken on the same subject over time are likely to be correlated. For the fixed part each regression coefficient is assumed to take the same fixed value for all people whereas the random effects are effects assumed to vary from person to person. Because treatment effect interpretation is easier for models with a linear predictor compared to models with a non-linear predictor the former are preferred.

Up to two longitudinal LME models will be fitted to the averaged mania scores with each model including the following covariates for the average response part of the model: time (as a continuous variable based on date the LIFE was completed relative to randomisation date), treatment arm along with the covariates considered for the Cox model fitted to the primary outcome (i.e., sex, number of previous episodes and wave). The baseline value for the mania score will be added. If this term is constant then this covariate will not be included in the model.

Based on available data, time will be centered to provide an estimate of the treatment effect and treatment by time effect.

We also specify a pair of correlated random effects: an intercept and linear slope where subjects have their own slope representing individual subjects' variations from the average slope. In addition, therapy group will be included as a random effect subject to model fitting constraints. Inclusion of therapy group takes account of group clustering effects. The models, in decreasing order of complexity, are as follows:

Model 1: As there may be a faster rate of recovery/decline in one group than the other a time with intervention arm interaction will be fitted. This main effects interaction between time and treatment arm will be tested for statistical significance.

Model 2: If the interaction in Model 1 is not significant then this term will be omitted and Model 2 will be fitted to test whether there is a systematic effect of treatment arm.

If there are convergence problems when fitting the models then the random slope and/or random therapy group terms will be omitted from the model. Restricted maximum likelihood will be used to fit the models.

Based on the final model the estimated treatment difference between the two groups, 95% confidence interval and P-values will be tabulated. This will be based on a wald test i.e. the estimated coefficient divided by the standard error of the coefficient.

The same process will be used to select and summarise a final model for the depression scores.

7.2.3 LME Inference and missing data

Of note, by using maximum likelihood for these models, "Missing At Random" is assumed for drop-out i.e., missing outcome data is conditional on observed data. Under this assumption it is assumed that future behaviour, given the past, is the same for all, whether a participant drops out or not. This allows distributional information to be "borrowed" from those who remain on the trial and applied to those who drop-out given they have the same covariate set up until the time of dropout. Therefore, the estimand of treatment effect is what would be seen if all participants had remained on the study until the end.

7.2.4 LME Model Diagnostics

The distributional assumptions of normality will be assessed at the time-point, subject and therapy group level. Where there is evidence of non-normality outcome data may be transformed. Particular observations that have unusually large influence on the results will be identified and the analysis repeated with them omitted.

7.2.5 Relapse rates

For each of the periods 1- 32, 33-64 and 65-96 weeks the number of people (percentage of randomised) who have had a mania relapse at least once during each of these intervals will be presented by arm. This will be repeated for depression and also for mania and depression combined. Only summary statistics without any inference will be presented as these analyses are similar to those based on time to event.

7.3 Survival models for Recurrence of Mania and Depression, separately

The survival analyses for the two secondary outcomes will follow Section 7.1 and will incorporate medical notes data, if appropriate.

7.4 Longitudinal Models for Other Continuous Secondary Outcome Measures

The approach will be essentially the same as in 7.2.2 using actual time from randomisation as a covariate.

8.1 Preferences and Adherence to Therapy

Participants are asked at baseline how effective they think each of the two treatments is likely to be and if they have a preference as to which group they are allocated. People were told that this wouldn't influence the randomisation to group in any way. This data will be summarised in the statistical report.

The total number of trial therapy sessions attended by each participant will be calculated from treatment arm attendance sheets. Summary statistics on the number of sessions attended will be tabulated by treatment arm . Baseline factors (see Section 6.1) that influence attendance of treatment will be investigated using a negative binomial model for count data. The reason for stopping therapy will also be tabulated by treatment arm where available.

8.2 Treatment Moderators

The pre-specified treatment moderators are number of previous episodes (<7, 8-19,20+; entered as a linear factor) and preference as to which group they are allocated. For each moderator, a treatment by baseline covariate interaction will be fitted to investigate how these covariates influence the treatment effect on the primary outcome, time to mania or depression relapse.

References

1. Colom F, Vieta E, Martinez-Aran A, Reinares M, Goikolea JM, Benabarre A, Torrent C, Comes M, Corbella B, Parramon G, Corominas J: A randomized trial on the efficacy of group psychoeducation in the prophylaxis of recurrences in bipolar patrients whose disease is in remission. *Arch Gen Psychiatry* 2003, **60:**402-407.
2. Colom F, Vieta E, Reinares M, Torrent C, Goikolea JM, Gastó C: Psychoeducation efficacy in bipolar disorders: beyond compliance enhancement. *J Clin Psychiatry* 2003, **64:**1101-1105.
3. Castle D, White C, Chamberlain J, Berk M, Berk L, Lauder S, Murray G, Schweitzer I, Piterman L, Gilbert M: Group-based psychosocial intervention for bipolar disorder: randomised controlled trial**.** *Br J Psychiatry* 2010, **196:**383-8.
4. D'Souza R, Piskulic D, Sundram S: A brief dyadic group based psychoeducation program improves relapse rates in recently remitted bipolar disorder: a pilot randomised controlled trial**.** *J Affect Disord* 2010, **1-3:**272-6.
5. White IR, Thompson SG. Adjusting for partially missing baseline measurements in randomized trials. Statistics in Medicine 2005; **24**: 993-1007.
6. Lobban F, Taylor L, Chandler C, Tyler E, Kinderman P, Kolamunnage-Dona R, et al. Enhanced relapse prevention for bipolar disorder by community mental health teams: cluster feasibility randomised trial. Br J Psychiatry 2010; **196:**59-63.

Appendix: Potential Additional Analyses

1. Treatment Mediators

Analysis of treatment mediators will depend on there being evidence of a treatment effect and will therefore be part of a later exploratory analysis. The effect of treatment on the mediators will be investigated separately from the main statistical analysis following the proposed discussion of causal pathways.

2. Exploratory Analyses

The following are potential additional analyses of the SCID LIFE scale for discussion:

- To take account of substance use disorder and personality disorder at baseline the survival and final LME models for mania and depression will be re-fitted additionally controlling for both these variables.
- Recovery from episode (a person who scored a 5 or 6, who score only 1 or 2 on both the SCID-LIFE mania or depression scales for 8 consecutive weeks.
- Remission, defined as those persons who have scored a 5 or 6 who then achieve a score of 4 or less on both SCID LIFE mania and depression scales for 8 consecutive weeks.
- Proportion of time spent in a significant mood state (score of 5 or 6); time spent in a mild but impairing mood state (3 or 4) time spent "well" (score of 1 or 2) {see below}
- Cycling
- An extended lme model with between and within reporting "diary" period random effects
- An extended lme model with a time by treatment interaction as a random effect term
- The total number of relapses a subject has over their follow-up period will be calculated. Poisson and negative binomial models for count data will be fitted to this data. The choice of model depends on goodness of fit. If there is evidence of overdispersion then the latter is preferred.
- A model for the overall proportion of time in relapse.
- The SCID LIFE scores will be grouped into 4 categories:1 or 2; 3 or 4; 5; 6. Compositional data plots showing the percentage of time (out of 100%) in each category over 16 weekly periods will be plotted.
- The number (percent) of times a subject goes from a 1 or 2 to a 3 or 4; 1 or 2 to a 5 or 6 ; 3 or 4 to a 5 or 6