STATISTICAL ANALYSIS PLAN

FOR

CALMED TRIAL

Date: 08/07/2021

Version: 2

North Wales Organisation for Randomised Trials in Health (NWORTH)

Bangor University, College of Health & Behavioural Sciences,

Institute of Medical & Social Care Research, Y Wern, Normal Site, Holyhead Road, Bangor, Gwynedd LL57 2PZ

Telephone: 01248 388095 Email: nworth@bangor.ac.uk Website: http://nworth-ctu.bangor.ac.uk/

The clinical effectiveness and cost effectiveness of clozapine for inpatients with borderline personality disorder: randomised controlled trial (CALMED).

IRAS Project ID:

REC:

Sponsor: Imperial College London

Funder: National Institute of for Health Research Health Technology Assessment Programme

EudraCT: 2018-002471-18

Based on protocol version: v9 (8th December 2020)

Chief Investigator: Mike Crawford

Trial Manager: Verity Leeson

Principal Statistician: Zoë Hoare

Trial Statistician: Rachel Evans

Approval signatures

Author

Rachel Evans	Signature:		Date: 08/07/2021
Senior Trial Statistician			

Approved by:

Zoë Hoare Principal Statistician	Signature:		Date: 09/07/2021
Mike Crawford Chief Investigator	Signature:		Date: 14/07/2021

DOCUMENT HISTORY

Updated version no.	Effective date	Authorship	Section changed	Summary/justification of changes
1	*n.a	Rachel Evans	n.a	New document
2	14/07/2021	Rachel Evans	Text on page 5 Table 2 - IPDE Appendix 1 – IPDE Appendix 2 added	Recruitment issues can't necessarily be attributed to negative social media IPDE scoring updated

*SAP 1 signed by RE and ZH. SAP updated to v2 prior to sign off by CI.

TABLE OF CONTENTS

1. ACRONYMS AND DEFINITION OF TERMS

Acronym	Meaning
AE	Adverse Event
ADSUS	Adult Service Use Schedule
ANCOVA	Analysis of Covariance
ANNSERS	The Antipsychotic Non-Neurological Side Effects Scale
AR	Adverse Reaction
BARS	Brief Adherence Rating Scale
BPD	Borderline Personality Disorder
BPRS	Brief Psychiatric Rating Scale
CACE	Complier Average Causal Effect
CI	Chief Investigator
CONSORT	Consolidated Standards of Reporting Trials
CRF	Case Report Form
CTIMP	Clinical Trial of Investigational Medicinal Product
CTU	Clinical Trials Unit
DMC	Data Monitoring Committee
Ecrf	Electronic Case Report Form
EDC	Electronic Data Capture
GCP	Good Clinical Practice
GLM	Generalised Linear Model
HEAP	Health Economics Analysis Plan
IMP	Investigational Medicinal Product
IPDE	International Personality Disorder Examination
ITT	Intention to Treat
MAR	Missing at Random
MCAR	Missing Completely at Random
MNAR	Missing Not at Random
MHRA	Medicines and Healthcare Products Regulatory Agency

Mitt	Modified Intention to Treat
MOAS	Modified Overt Aggression Scale
NHS	National Health Service
NIHR	National Institute for Health Research
NWORTH	North Wales Organisation for Randomised Trials in Health
REC	Research Ethics Committee
SAE	Serious Adverse Event
SAP	Statistical Analysis Plan
SAR	Serious Adverse Reaction
SASPD	Standardised Assessment of Severity of Personality Disorder
SCID-II	Structured Clinical Interview for Axis II
SD	Standard Deviation
SOP	Standard Operating Procedure
SUSAR	Suspected Unexpected Serious Adverse Reaction
TMG	Trial Management Group
TSC	Trial Steering Committee
ZAN-BPD	Zanarini Rating Scale For Borderline Personality Disorder

2. STATISTICAL ANALYSIS PLAN AUTHORSHIP

This SAP has been authored by Rachel Evans (Senior Statistician NWORTH CTU), Trial Statistician for the CALMED Trial, with input from Zoë Hoare, Senior Trial Statistician for the trial (Principal Statistician NWORTH CTU), Mike Crawford (Chief Investigator) and Verity Leeson (Trial Manager).

The study began recruitment in September 2019 however, due to several difficulties such as recruiting in this population and the impacts of COVID-19, which are still ongoing, the study struggled to reach recruitment targets. Efforts were made to increase recruitment to the study but unfortunately recruitment to the study remained an issue.

The sample size of the study was re-calculated and reduced, see section 4.1, however in March 2021, with agreement from independent committees and the funder, the study was terminated. At the time of termination, a final sample of 29 randomised patients was reached.

This document details the planned analysis according to the protocol, given the full sample, and the analysis that will be carried out for the study given the sample reached. It should be noted that the sample is very small, and far below the statistical power required to evaluate the intended research question. Therefore, all results presented following the analysis should be interpreted with caution.

3. INTRODUCTION

3.1 BACKGROUND AND DESIGN

Inpatients with severe borderline personality disorder experience high levels of emotional distress which may be accompanied by high levels of self-harming and aggressive behaviour. Open label studies of treating inpatients with clozapine show major improvements in mental health and reductions in self-harming behaviour. However, to date there are no published or on-going clinical trials of clozapine for people with borderline personality disorder. While clozapine has the potential to generate considerable savings for the NHS, the true risks and benefits of treating people with clozapine are unknown. The CALMED trial is a fully powered two-arm, double-blind, placebo-controlled randomised trial which is designed to establish the clinical effectiveness and cost effectiveness of clozapine for inpatients with severe borderline personality disorder.

Design & Methodology

A two-arm, parallel-group, placebo-controlled trial with an internal pilot and an integrated economic evaluation. The study is funded by the National Institute of Health Research (NIHR) Health Technology Assessment programme. Study groups are standard care plus clozapine, titrated up to a maximum dose of 400mg daily, versus standard care plus an inert placebo. Standard care will include access to psychological interventions, occupational activities and nursing care that are provided to inpatients with BPD.

222 inpatients with BPD will be recruited from seven centres in England over a 24-month period. All participants will be followed-up at three and six months using a battery of assessment scales to measure mental health (Zanarini rating scale for Borderline Personality Disorder - ZAN-BPD, and the Brief Psychiatric rating Scale - BPRS), incidence of violence to self or others (MOAS), acts of deliberate self-harm, health-related quality of life (EQ-5D-3L), side effects of treatment, adherence and adverse reactions. Resource use and costs will be assessed using a modified version of the Adult Service Use Schedule and with nationally available unit costs. We will continue to collect economic data from clinical records on use and type of inpatient treatment, use of community services and quantity and type of medication that participants take throughout the study. We will use a flexible dosing regimen of clozapine. Dosing will start with 12.5 mg once daily and be titrated to 300mg over a 15-day period. Study participants may be prescribed a dose of up to 400mg of clozapine daily, depending on clinical response, patient preference and side effects. The dose may be

maintained at or reduced to a lower dose at any time. Equivalent numbers of placebo capsules will be administered to participants in the control arm of the trial. All those taking part in the study will continue to receive all other treatments as usual.

The trial will involve two linked phases:
Phase 1 - An internal pilot in which participants will be recruited from all centres over a six-month period. Data from the internal pilot will be presented to the Trial Steering Committee (TSC) indexed against a priori stop/go criteria, that will be used to determine if the study progresses to phase 2.
Phase 2 - Full trial across all sites over a further 18-month period.

Progression criteria will be assessed at six months according to the following three parameters:

- Number of participants randomised in the first six months. Our target is 55. If above 42 (75% of the target) recruitment would continue. If between 35 and 41 we would discuss continuation with the independent committees and funder. If below 32 (less than 60% of the target) we would stop. Mitigating circumstances for under-recruitment such as a delay in opening one or more of the pilot sites will be discussed with the TSC and may result in an adjustment to the participant target or pilot period.
- At least 75% of the participants recruited to the active arm of the trial in the first three months of the pilot phase of the trial will have started study medication within four weeks of randomisation.
- Three-month follow-up data will have been collected from >70% of those randomised during the first three months of the pilot study.

3.2 TRIAL OBJECTIVES

Primary Objective:

For people receiving inpatient treatment for severe borderline personality disorder who have not made an adequate clinical response despite receiving usual care (including at least three months taking another antipsychotic drug), does the addition of clozapine to their usual care lead to improved mental health six months later, compared to adding an inert placebo to their usual care?

Secondary Objectives:

i) For people receiving inpatient treatment for borderline personality disorder who have not made an adequate clinical response to usual care, does the addition of clozapine lead to greater improvements in health-related quality of life, greater reductions in aggression, suicidal behaviour, and use of services compared to adding a placebo?

ii) For people receiving inpatient treatment for borderline personality disorder who have not made an adequate clinical response to usual care, does the addition of clozapine provide a cost-effective treatment compared to the addition of a placebo?

3.3 CONSORT DIAGRAM

CONSORT 2010 Flow Diagram

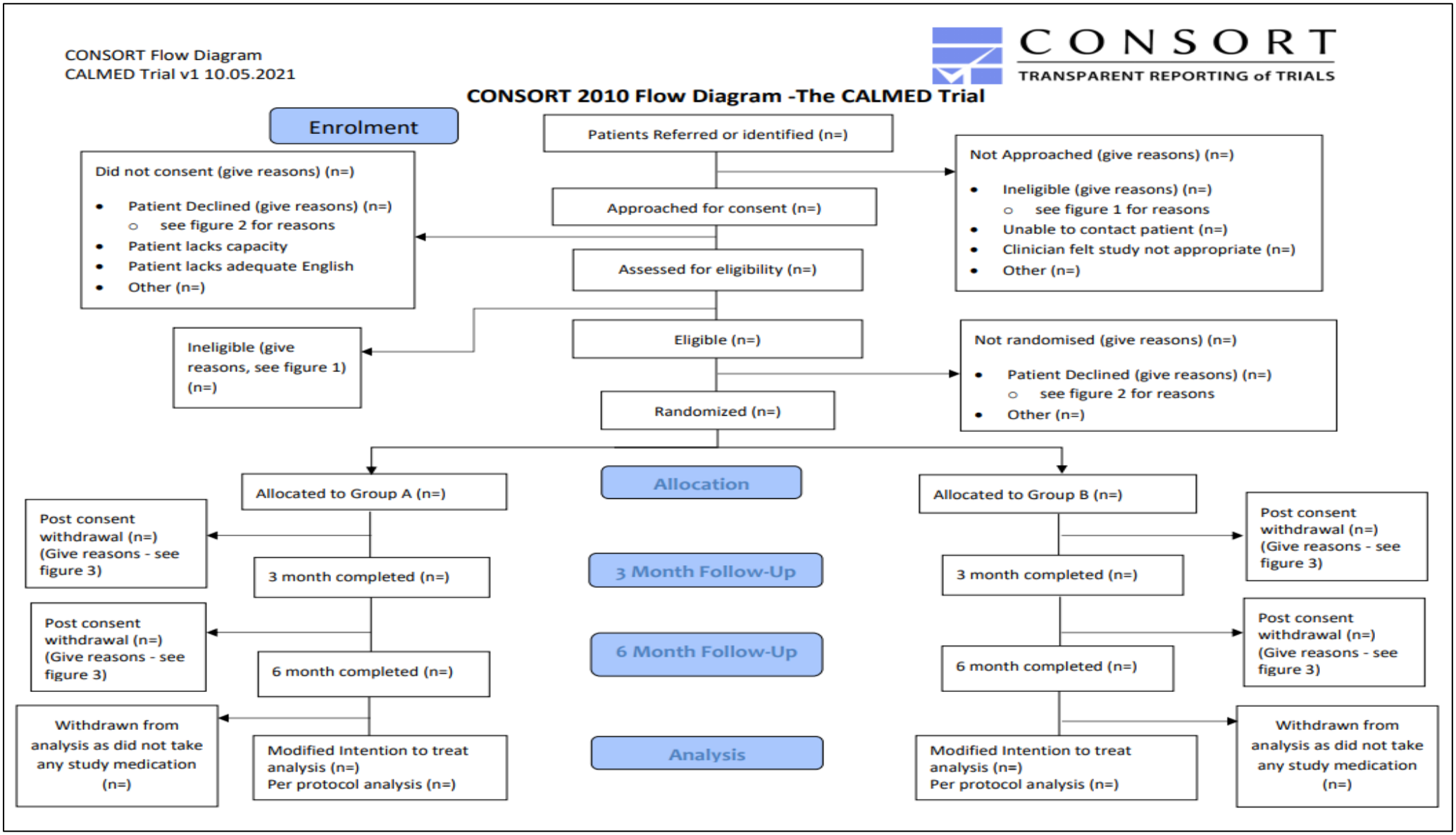

Figure 1: CALMED study Flow Chart for CONSORT figures

CONSORT Flow Diagram
CALMED Trial v1 10.05.2021

CONSORT
TRANSPARENT REPORTING of TRIALS

Reasons for Ineligibility

Inclusion criteria:

- Aged 18 years or over
- Currently an inpatient on a mental health unit
- Meeting DSM-IV diagnostic criteria for borderline personality disorder
- Failure to make an adequate clinical response to taking antipsychotic medication other than clozapine for at least three months
- Have a satisfactory pre-treatment full blood count
- Have had their weight and blood glucose recoded in their clinical records
- Has severe personality disorder

Exclusion criteria:

- Current clinical diagnosis of schizophrenia or bipolar I disorder
- Prescribed clozapine within the previous two weeks
- Is pregnant or trying to conceive, breastfeeding, or a woman of childbearing potential not using a highly effective birth control.
- Contraindications to clozapine
- Due to be discharged from the unit within the following two weeks and it is not possible to continue the necessary monitoring of physical health as an outpatient
- Unable to speak sufficient English to complete the baseline assessment
- Unwilling or unable to provide written informed consent to take part in the study
- unable to undertake regular blood tests

Figure 1: List of reasons for ineligibility

Reasons patient declined

- Does not want to be randomised / risk placebo
- Does not want to take study medication
- Does not want to complete questionnaires/interviews
- Does not want to have blood tests
- Other

Figure 2: List of reasons for patient declining participation

Reasons for post-consent withdrawal

- Protocol Deviation
- Adverse Event
- Death
- Withdrawal of Consent
- Lost to Follow Up
- Trial terminated by sponsor
- Other

Figure 3: List of reasons for post-consent withdrawal

Figure 2: CALMED study Flow Chart for CONSORT figures - figure references

4. STATISTICAL CONSIDERATIONS

4.1 SAMPLE SIZE JUSTIFICATION

The sample size calculation for the study was based on the primary hypothesis that, for inpatients with borderline personality disorder, the addition of clozapine to usual treatment reduces symptoms of the disorder measured at six months (standard deviation 7.89) using the Zanarini Rating scale for Borderline Personality Disorder (ZAN-BPD). 166 participants (83 receiving clozapine and 83 receiving placebo) will need to be randomised to have 90% power to detect a four point clinically important difference in ZAN-BPD score at six months, using a 0.05 level of statistical significance. To take account of 25% loss to follow-up we planned to recruit 222 subjects. We will seek to further increase the sample size to ensure that we have data on 166 participants who took at least one dose of trial medication.

The above original sample size estimate was calculated with a two-sample t-test method, this is used with the understanding that the model used for analysis (i.e. ANCOVA) will have higher statistical power which is seen as a conservative approach and often adopted to power studies when there is little additional data available. An alternative approach would be to use an ANCOVA model calculation. This method requires an additional assumption of the correlation covariate for the calculation, which there is often no robust data available for, thus reducing the certainty in the sample size proposed. However, based on ZAN-BPD data held by the study team and available in the literature we felt confident to estimate the correlation co-efficient for the current study. Some caution is still being applied by taking the most conservative estimate of those available.

Using data from a previous trial (Labile) and the current CALMED data the correlations were calculated were between 0.2 and 0.4. Walters et. al. (2019) found in most cases the correlation between baseline and follow up assessments was between 0.4 and 0.6. It is possible that a correlation of 0.4 could be seen; however, assuming the correlation to be higher than it is would reduce the potential power of the final analysis. Therefore, we propose to take a more conservative approach and use an estimate of 0.3.

It was recommended that the trial aims to recruit, for analysis, a sample of N=118. Which, with a 0.3 correlation of baseline and follow-up scores, would give 90% power to detect a 4.0 difference on the ZAN-BPD scale with a 7.89 SD at a 5% significance level.

By adopting this alternative method of sample size estimation additional information can be taken into account ultimately relating the sample more closely to the analysis method that will be used. A benefit of this alternate is that all original parameters will remain the same (mean difference of 4.0 and assumed standard deviation of 7.89) thus preserving the primary research aim.

4.2 RANDOMISATION

Randomisation was via a secure online system using a sequentially randomised dynamic adaptive algorithm stratified by centre (list), ward type (general adult, low secure, medium secure and high secure) and gender (male or female), hosted by NWORTH, Bangor university. Within the algorithm, the likelihood of the participant being allocated to each treatment group is recalculated based on the participants already recruited and allocated (Russell et. al., 2011). This recalculation is done at the overall allocation level, within stratification variables and within stratum level (the relevant combination of stratification levels). By undertaking this re-calculation, the algorithm ensures that balance is maintained within acceptable limits of the assigned allocation ratio while maintaining unpredictability.

Randomisation creates a unique randomisation code for the participant. A trial prescription form will then be completed by the local investigator. This will be sent to the pharmacy that will dispense the trial medication to the ward where the participant is being treated.

4.3 LEVELS OF CONFIDENCE AND P-VALUES

All statistical tests for the primary endpoints will be two-sided and will be performed using a 5% significance level. Confidence intervals for estimated effects will be presented as 95% and two-sided. As this study will not provide conclusive evidence and will lack statistical power, secondary outcomes will not be adjusted for multiple comparisons. Results will be interpreted with caution, 95% confidence intervals will be reported alongside any p-value or effect.

4.4 ADHERENCE AND COMPLIANCE

The planed analysis for Trial Adherence was to investigate the effect of treatment adherence using complier average causal effect (CACE) estimation methods. Intention To Treat (ITT) analysis does not represent treatment effect under non-compliance of treatment, we therefore planned to use CACE analysis to explore whether treatment effect is directly affected by the level of compliance.

However, due to the small sample size the above analysis is not appropriate, therefore, compliance and adherence will be reported descriptively as both a dichotomous and continuous variable. The dichotomous: whether the participant took medication at a dose of 100mg or more without interruption during the six months prior to the final follow-up interview; b) Continuous - percentage of medication the participant took in the month prior to competing the six month follow up assessment using data from the Brief Adherence Rating Scale.

This will be presented overall and split by treatment group, continuous variables will be reported with mean values, standard deviations and ranges. Categorical variables will be presented with counts and percentages.

4.5 PROTOCOL VIOLATIONS

A record is kept of any action that differs from that described in the protocol, study or sponsor procedures that is not approved by the sponsor/REC/MHRA prior to its implementation. Non-serious examples are considered deviations and include issues such as a study visit date being outside the window defined in the protocol. More serious or systematic non-compliance is considered a violation. Any protocol violations or deviations will be reported descriptively, and sensitivity analysis will not be conducted on these. Per protocol analysis is however planned for those who took medication to at least 3-months (see section 4.8 and 6.10).

4.6 MISSING DATA

Missing observations are expected within the dataset. Completion rates of outcome measures will be summarised in the final analysis report, at the item and overall level.

Missing data – outcome item level

For missing items within an outcome measure, the published rules for completing missing data for the relevant measure will be applied. Where there are no missing data rules for the measure, if the number of missing items on an outcome measure is 20% or less, then the missing value for the item will be substituted by the individual's mean score for the remaining items on the scale (Bono, Ried, Kimberlin and Vogel 2007). If there are more than 20% missing items in the scale the outcome measure will not be calculated for the participant at that time point.

Missing data –outcome total score level

A missing completely at random test devised by Little (1998) will be performed to assess whether the data is missing completely at random (MCAR). If this test indicates that data is missing completely at random than analyses will be based on complete cases, else independent t-tests will be conducted to investigate whether the data is missing at random (MAR). If these tests suggest that the missing data is MAR, then predictive mean matching multiple imputation method will be employed. Otherwise, additional modelling guided by clinical knowledge would be required to simulate the missing data mechanism and impute the missing data.

Analysis will be conducted using multiple imputation for the primary and secondary analysis outcome measures. Complete case analysis will be conducted as sensitivity analysis (see section 6.10) to check the impacts that missing data may have had on the results, this is especially important given the small sample and high withdrawal (in relation to the sample) rate.

For imputation, predictive mean matching multiple imputation methods will be adopted. For multiple imputations, the number of imputations completed is dependent upon the percentage of missing data (White, Royston, & Wood, 2011). The current dataset has 18% of the primary outcome missing, which would require 18 imputation sets, which is very large given the total study sample, therefore, the number of imputations with this sample will be limited to m=5, White (2011) state that "Standard texts on multiple imputation suggest that small numbers of imputed data sets (m =3 or 5) are adequate".

The missing outcome measures will be imputed using group allocation, stratification variables, outcome measure baseline scores and any other baseline characteristics that are deemed to be predictors of missingness. Baseline characteristics will be assessed for being predictors of missingness by running statistical tests on completers versus non completers and evaluating if any differences are present between the two. Baseline factors to be assessed as predictors of missingness include; sex at birth, centre, ward type, ethnicity, Gender (self-identified) and baseline ZAN-BPD scores.

4.7 ASSUMPTION CHECKING

All assumptions relating to the GLM models will be checked and evaluated whether appropriate to use with the data. Table 1 contains details of the assumptions associated with a GLM model and the methods to be used to assess these assumptions.

Table 1: List of assumptions and checks for a GLM

Assumption	**Checking**
Generalised Linear mixed model	
Linearity - the relationship between the independent and dependent variables to be linear	scatter plots of the model residuals vs predictor
Residuals/errors are independent - Little or no autocorrelation in the data. (residuals should be independent from each other)	Scatter plot Durbin-Watson test
Residuals/Errors are normally distributed	Q-Q-Plot
Residuals/Errors have constant variance - There should be no homoscedasticity of error terms	Scatter plot of standardized residuals versus predicted values
No or little multi-collinearity - (independent variables should not be highly correlated with each other)	inspection of correlation coefficients and Tolerance/VIF values

Outliers or unusual values will be assessed by running Grubbs (1969) test for outliers and visually inspecting a boxplot. No outliers will be discarded if they are within plausible range. Primary analysis will be conducted keeping the outliers in the dataset and if necessary, sensitivity analysis will be conducted by removing the outliers and evaluating any effects on the results and conclusions of analysis. Any outliers removed will be fully reported.

The distribution of the data will be checked and based on this a decision will be made as to whether a transformation should be applied and if so, which transformation should be used. If transformation is required, the distribution of the transformed data will be checked. Analysis will be reported on the original scale, transforming variables back. If a transformation is inappropriate, then nonparametric analysis methods will be considered.

4.8 ANALYSIS SETS

All primary analysis will be using modified ITT analysis (mITT), in which all randomised participants will be included who have taken at least one dose of trial medication.

For participants that were randomised but did not take any trial medication, they will be included in summary descriptive statistics only as they are not part of the mITT dataset (see figure 1).

A per protocol analysis set will also be used. This will consist of participants who took trial medication, at least until the first follow-up at 3 months. i.e. all those in the mITT set who withdrew from trial medication prior to their 3 month follow up are removed in addition to the participants randomised that didn't take trial medication at all.

5. DATA

All data will be handled in accordance with General Data Protection Regulation (GDPR, 2018). The Case Report Forms (CRFs) will not bear the subject's name or other personal identifiable data.

All electronic databases will use a patient identification number rather than the patient's name. Hard copies of data sheets linking the patient identification number to the person's contact details will be kept securely in a locked filing cabinet in a locked office and will only be accessible to a small number of people who are involved in the study. The database will be stored on a network drive at University of Bangor, which is backed-up daily.

Data monitoring will be carried out according to the trial-specific monitoring plan. At the end of the trial, data monitoring and cleaning will be completed and the database "locked" so that no further data entry is possible. At this point the data will extracted by the trial statistician along with the randomisation list stating whether participants were allocated to Arm A or B. The data will be analysed without knowledge of which arm relates to the active trial medication. Any subsequent 'unblind analysis' will be carried out once all blinded analysis is complete, reported to the study team and the Trial Statistician has been unblinded officially in accordance with NWORTH SOP.

5.1 TIME POINTS OF OUTCOMES MEASURES

Table 2 below details a list of the outcome measures being collected in the study and at which time points.

Table 2: Details of measures being collected for study

Assessments	Screening	Baseline	3-month	6-month	12 and 18-month	Analysis
Structured Clinical Interview for Axis II Personality Disorders (SCID-II)	X	-	-	-	-	Screening tool only - not being presented or analysed
Psychosis Screening Questionnaire	X	-	-	-	-	Screening tool only - not being presented or analysed
International Personality Disorder Examination (IPDE)	X	-	-	-	-	Screening tool only - being presented descriptively but not being analysed
Standardised Assessment of Severity of Personality Disorder (SASPD)	-	X	-	X	-	GLM analysis at 6 months
Zanarini rating scale for Borderline Personality Disorder (ZAN-BPD)	-	X	X	*X	-	GLM analysis at 3 and 6 months
Brief Psychiatric Rating Scale (BPRS)	-	X	X	X	-	GLM analysis at 3 and 6 months
Acts of Deliberate Self Harm Inventory	-	X	X	X	-	To be presented descriptively (count data) at each time point
Modified Overt-Aggression Scale (MOAS)	-	X	X	X	-	GLM analysis at 3 and 6 months
EuroQoL EQ-5D 3 level	-	X	X	X	-	For health economic analysis (being conducted as separate analysis)
The Antipsychotic Non-Neurological Side Effects Scale (ANNSERS)	-	X	X	X	-	GLM analysis at 3 and 6 months
The Extrapyramidal Side Effects Scale	-	X	X	X		GLM analysis at 3 and 6 months
Brief Adherence Rating Scale (BARS)	-		X	X	-	T-test at 3 and 6 months
Adult Service Use Schedule (ADSUS) Use of inpatient and community mental health services	-	X	X	X	X	For health economic analysis (being conducted as separate analysis)
Trial Arm allocation guess				X		Independent variable for sensitivity analysis - not being used as an outcome measure.

*Primary outcome is Zan-BPD at 6 months.

5.2 DEFINITIONS AND CALCULATIONS OF OUTCOME MEASURES

The below list indicates the primary and secondary outcomes for the study.

Primary outcome

1. The primary outcome is total score on the Zanarini rating scale for Borderline Personality Disord er (ZAN-BPD) at six months.

Secondary outcomes

2. Total score on the Zanarini rating scale for Borderline Personality Disorder at three months.
3. General mental health using the Brief Psychiatric Rating Scale (BPRS) at three and six months.
4. Incidence and severity of suicidal behaviour using the Acts of Deliberate Self-Harm Inventory.
5. Level of aggressive behaviour using the Modified Overt Aggression Scale
6. Health related quality of life using the EQ-5D-5L.
7. Side effects of medication using the Antipsychotic Non-Neurological Side Effects Scale (ANNSERS) and motor and extrapyramidal side effects using the Extrapyramidal Side Effects Scale.
8. Incidence of withdrawal of trial medication due to adverse effects.
9. Medication adherence at three and six months using the Brief Adherence Rating Scale.
10. Resource use collected using a modified version of the Adult Service Use Schedule and by examining clinical records at six, 12 and 18 months, This will include detailed information about length of inpatient treatment and type of ward (high, medium, low secure, Psychiatric Intensive Care, general adult etc.), contacts with community mental health services and emergency medical services, and the type and dose of psychotropic medication that people are prescribed.

For each measure, its associated analysis to be conducted is listed in table 2. Appendix 1 contains details and definitions of scoring these measures where applicable.

As detailed in the table, the SCID-II, Psychosis screening questionnaire and the IPDE are measures all used for screening and will not analysed. The IPDE will be used to describe the personality profile of the participants. The EQ-5D-3L and ADSUS are being collected for health economic analysis, which will be analysed and reported separately and summarised in a separate health economics analysis plan (HEAP). Additionally, physical health measures for participants are collected on a weekly basis as part

of physical and safety monitoring, these measures will not be summarised or presented as part of the analysis covered in this SAP, we will report on adverse events (see section 5.3 below).

5.3 SAFETY DATA

Safety data, covering the number of events and number of subjects affected split by treatment arm and categorised by event type (I.e. AE, AR, UAR, SAE, SAR, SUSAR) will be presented. Frequencies of system organ class (MEDDRAA code) will also be presented split by treatment arm.

Patients will be categorised based on whether they had an event or not (for each type) and this data will be summarised overall and by group. If appropriate (i.e. enough representation of yes/no) chi square test will be run on these and allocation group.

5.4 UNBLINDING THE TRIAL

Once the analysis detailed in this SAP has been executed procedures for unblinding the trial statistician will follow NWORTH SOP 5.03. Once the results of the analysis have been reported to the TMG and with approval from NWORTH Principal Statistician, the trial statistician will request unblinding from NWORTH IT. The CI of the trial will be informed that the trial statistician is to be unblinded. In line with SOP 5.03, a form will be completed and sent to NWORTH IT who will then inform the statistician of the group allocation details.

6. STATISTICAL ANALYSES

6.1 ANALYSIS TIME FRAME

Table 3: Planned target timeline for analysis dependent on recruitment period

Task	Planned time frame in line with study closeout
Last patient recruited	January 2021
Final patient visit	July 2021
*Database Lock	~August 2021
*Analysis	~September 2021

**Database lock will occur 30 days following final patient visit, given data cleaning has been finalised and all queries dealt with. Analysis report will be finalised and sent to the study team within one month from data lock.*

6.2 BLIND REVIEW OF THE DATA

The outcomes for this study are continuous measures and we would not envisage any major issues that would require a deviation from analysis. The assumptions for the models will be checked as indicated in section 4.7. Given this and the nature of the analysis, we will not be doing a blind review of the data. Any deviations from the intended analysis along with any transformations or assumptions violated will be described and fully justified in the write up of the analysis report.

6.3 BASELINE ANALYSIS

Demographics and patient characteristics collected at baseline will be presented descriptively both overall and by treatment arm. No statistical testing of baseline differences will be completed, only to visually assess balance (de Boer et al., 2015; Moher, 2010). Continuous variables will be reported with mean values, standard deviations and ranges. Categorical variables will be presented with counts and percentages. If data are not normally distributed, then medians and interquartile ranges will be reported.

6.4 INTERIM ANALYSIS

There are no planned interim analyses and as such the study is not powered to achieve these.

6.5 CONSORT ANALYSIS

The patient flow information, shown in section 3.3 as advised by CONSORT reporting standards (Schulz et al., 2010), will be completed with values relating to participants numbers. Data will be presented on screening, eligibility, recruitment, treatment discontinuation, study withdrawals and lost to follow up, presented for the entire study and split by recruitment site. Where possible, reasons for ineligibility, non-recruitment, treatment discontinuation and withdrawal will be reported. From these data related eligibility, recruitment and retention rates will be calculated and presented.

6.6 DESCRIPTIVE STATISTICS

Descriptive statistics of the data will be presented in the final analysis report. This will include randomisation figures, demographics, other data characteristics of the outcome measures and data completeness levels. All will be presented overall and split by allocation group.

Figures will be presented for the Randomisation data by each stratification variable;

- Allocation group
- Centre

- Gender (Sex at birth)
- Type of ward

The following demographics and data characteristics will be presented;

- Ethnicity
- Gender (self-identified)
- Age
- ICD-10 codes current admission and previous admission
- IPDE scores

Descriptive statistics of the primary outcome measures will be produced, these will be presented overall and split by allocation group. Additionally, the primary outcome descriptive statistics at baseline will be split by ward type. Furthermore, 3 and 6 months by trial arm allocation perception, see section 6.10 for more details. Descriptive statistics of the secondary outcome variables as listed in table 2will also be presented (overall and split by group).

For all statistics, continuous variables will be reported with mean values, standard deviations and ranges. Categorical variables will be presented with counts and percentages. If data are not normally distributed, then medians and interquartile ranges will be reported.

6.7 ANALYSIS OF PRIMARY OUTCOME

The intention for the study was to analyse the primary outcome using model general linear model fitted at six months and adjusted for baseline score, allocation group and randomisation stratification variables (centre, gender and ward type). Additional covariates would have also been considered.

Due to the small sample, the primary outcome will be analysed, with a general linear model at six months adjusted for baseline score only along with site as a random effect. With such little data including all the intended variables (randomisation stratification variables) in the model would likely result in overfitting the data. Thus, this model in its most simple form seems the most reasonable analysis to assess the research question given the small sample. The study will not be statistically powered, and results will be interpreted with caution.

Primary analysis will be with the mITT dataset. However, further exploratory analysis consisting of a per protocol analysis will be conducted on the primary outcome as indicated in section 6.10.

6.8 ANALYSIS OF SECONDARY OUTCOMES

All secondary outcome measures are continuous measures and will be analysed as the primary outcome model with a GLM to assess the differences between the means of the treatment groups. The baseline measure scores of the associated outcome will be included in the models as covariates to account for the baseline scores. The BARS measure was not applicable at baseline, therefore independent t-test models will be run on that measure at 3 and 6 months.

Secondary outcomes include the ZAN-BPD at 3 months, and all other secondary measures (listed in table 2) at 3 and 6 months (other than the screening tools and health economic measures).

6.9 SUBGROUP ANALYSES

Subgroup analyses would have been considered if the trial had of reached the recruitment target, however with such little data subgroup analysis would be inappropriate. Descriptive statistics of the primary outcome variable will be presented by the below subgroups:

- Ward type
- Allocation treatment 'guess' (see section 6.10 below)
- Co-existing clinical diagnosis of antisocial personality disorder

6.10 SENSITIVITY ANALYSES OR MODEL TESTING

Some exploratory sensitivity analysis will be conducted; however, this will be limited due to the small sample. Sensitivity analysis results and interpretation, as with the main analysis, should be treated with caution due to the small sample and lack of statistical power.

Sensitivity analysis may be required around any assumptions made for the primary analysis (e.g. outliers, data distributions) but only if necessary. Complete case analysis will be conducted on the primary and secondary data models to check the impacts of multiple imputation on the results at 6 and 12 months.

As indicated, primary and secondary outcomes will firstly be analysed as a mITT analysis. Sensitivity analysis will be carried out on a per protocol analysis (participants who took trial medication, at least until the first follow-up at 3 months) at 6 and 12 months on the primary outcome.

A sensitivity analysis around treatment arm allocation perception will be conducted. This separates into two parts (blinded analysis and unblinded analysis). Firstly, blinded, we will look at the 'view of allocation' i.e. did their view of whether they had placebo or active impact on Zan-BPD scores. For both the assessor and patient, ZAN-PBD scores, will be summarised descriptively, split by 'view of allocation arm' (regardless of what they actually had). For participants the primary outcome model will be run with this variable included to assess any impacts on the scores. In addition, patients view of their allocation and whether they discontinued from study medication, and if so at what time point, will be presented descriptively. This will be conducted on the mITT data set and the per protocol set.

Subsequently, following all blinded analysis being conducted, statistics of numbers of patients and assessors that correctly guessed their allocation will be presented, and in which direction (i.e. thought they were on placebo and were right or wrong or thought they were on active and were right or wrong).

Finally, a sensitivity analysis that adjusts any difference seen between treatment arms according to whether the participant reported 'lethargy/lassitude', as measured by the ANNSERS, will be conducted. This will be carried out by including the ANNSERS item for 'lethargy/lassitude' as an independent variable in the analysis model on the ZAN-BPD at 3 and 6 months. This will be on both the mITT and per protocol analysis sets.

Descriptive statistics will also be presented for the number of participants who discontinued trial medication and were then prescribed Clozapine while continuing in the trial. These will be summarised as counts and percentages presented overall and by allocation group.

6.11 EXPLORATORY ANALYSES

The trial protocol indicates that "exploratory analysis of the potential modification of the treatment effect will also be undertaken for pre-planned variables of gender and baseline score on the BPRS. Initially an interaction term between each variable and allocated group will be added to the model. Further modelling will then be undertaken if appropriate." This will not be conducted due to the small sample.

Due to the COVID-19 pandemic Zan-BPD outcomes were collected via telephone rather than face to face as originally planned. The measure was scored but two individual researchers independently from another. The level of agreement between researcher on these scores will be assessed exploratively.

6.12 ECONOMIC ANALYSES

Health economic analysis will be conducted separately to the study analysis by the trial Health economist. Details on this analysis is to be included in a separate health economics analysis plan (HEAP).

6.13 SAFETY AND TOLERABILITY ANALYSIS

Safety and tolerability analysis is not applicable for the CALMED Trial and will not be conducted as part of the study analysis. Adverse event data for the trial will be summarised (see section 5.3) but not in relation to safety and tolerability analyses.

7. SOFTWARE

All quantitative analysis will be completed using STATA version 15 or later.

8. REFERENCES

Bono, C.A. , Ried, L.D. , Kimberlin, C.A. , & Vogel, W.B. (2007). Missing data on the Center for Epidemiological Studies Depression Scale: A comparison of four imputation techniques. Research in Social & Administrative Pharmacy, 3, 1-27.

David Moher, Sally Hopewell, Kenneth F. Schulz, Victor Montori, Peter C. Gøtzsche, P.J. Devereaux, Diana Elbourne, Matthias Egger, Douglas G. Altman, CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials, Journal of Clinical Epidemiology, Volume 63, Issue 8, 2010, Pages e1-e37, ISSN 0895-4356,

de Boer, M. R., Waterlander, W. E., Kuijper, L. D., Steenhuis, I. H., & Twisk, J. W. (2015). Testing for baseline differences in randomized controlled trials: An unhealthy research behavior that is hard to eradicate. International Journal of Behavioral Nutrition and Physical Activity, 12(1), 4. https://doi.org/10.1186/s12966-015-0162-z

GDPR, 2018. Guide to the General Data Protection Regulation. [online] GOV.UK. Available at: <https://www.gov.uk/government/publications/guide-to-the-general-data-protection-regulation> [Accessed 18 May 2021].

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. Technometrics, 11(1), 1-21. https://doi.org/10.1016/j.jclinepi.2010.03.004.

https://www.labile.org/

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. Journal of the American Statistical Association, 83(404), 1198-1202.

PASS [Computer software]. (2015). Retrieved from http://www.ncss.com/software/pass/

Russell D, Hoare ZSJ, Whitaker R, Whitaker CJ, Russell IT. Generalised method for adaptive randomisation in clinical trials. Statistics in Medicine 2011;30:922-34.

Schulz, K. F., Altman, D. G., Moher, D., & the CONSORT Group. (2010). CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. Trials, 11(1), 32. https://doi.org/10.1186/1745-6215-11-32

StataCorp. 2017. Stata Statistical Software: Release 15. College Station, TX: StataCorp LLC.

Walters SJ, Jacques RM, Dos Anjos Henriques-Cadby IB, Candlish J, Totton N, Xian MTS. Sample size estimation for randomised controlled trials with repeated assessment of patient-reported outcomes: what correlation between baseline and follow-up outcomes should we assume? Trials. 2019 Sep 13;20(1):566. doi: 10.1186/s13063-019-3671-2. Erratum in: Trials. 2019 Oct 28;20(1):611. PMID: 31519202; PMCID: PMC6743178.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. Statistics in medicine, 30(4), 377-399.

APPENDICES

APPENDIX 1 – OUTCOME MEASURES SUMMARY

Purpose	Scoring	Item Coding	Subscales	Direction	Missing value rules	Thresholds
Acts of Deliberate Self Harm Inventory						
Assess deliberate self-harm	Measure used in study is different from published measure. Scoring not applicable, descriptive statistics of events will be summarised for measure, like adverse event data presentation.	n/a	n/a	n/a	n/a	n/a
The Antipsychotic Non-Neurological Side Effects Scale (ANNSERS)						
Comprehensive measure for rating non-neurological adverse drug reactions (ADRs) to antipsychotic	43-item scale. No. of items for each side effect severity 'class' (mild, moderate, severe) is summed. (*Note item 44 is not included in total*) Total score is calculated by multiplying the number of items tallied for each severity class by the item code for that class. Total score ranges from 0 - 129 Example: None – 30, Mild total – 3, Moderate, total – 9, Severe – 1 Total (0x30) + (3x1) + (9x2) + (1x3) = 21	None – 0 Mild – 1 Moderate – 2 Severe - 3	none	Higher scores indicate more severe side effects i.e. "Higher worse"	None found	None found
Brief Adherence Rating Scale (BARS) *Matthew J. Byerly, Paul A. Nakonezny, A. John Rush (2008). The Brief Adherence Rating Scale (BARS) validated against electronic monitoring in assessing the antipsychotic medication adherence of outpatients with schizophrenia and schizoaffective disorder, Schizophrenia Research, Volume 100, Issues 1–3, Pages 60-69, ISSN 0920-9964, https://doi.org/10.1016/j.schres.2007.12.470.*						
Assess the oral antipsychotic medication adherence	4-item scale. 3 questions and one visual analogue scale (VAS) 1 rating (0-100%). The VAS rating serves as the final adherence determination, therefore, total score ranges from 0-100	n/a	None	Higher scores indicate better adherence i.e. "Higher better"	None found	None found

Brief Psychiatric Rating Scale (BPRS) *Zanello, A., Berthoud, L., Ventura, J., & Merlo, M. C. G. (2013). The Brief Psychiatric Rating Scale (version 4.0) factorial structure and its sensitivity in the treatment of outpatients with unipolar depression. Psychiatry Research, 210(2), 626–633. https://doi.org/10.1016/j.psychres.2013.07.001*						
measure patient's psychopathology	24-item scale, each rated on a 7-point Likert scale (see item coding). Total score is calculated by summing items, hence ranges from 24 – 168.	0 = not assessed (Missing) [not counted in final score]. 1 = not present, 2 = very mild, 3 = mild, 4 = moderate, 5 = moderately severe, 6 = severe, 7 = extremely severe	None found	Higher scores indicate more severe symptoms i.e. "Higher worse"	None found	None found
The Extrapyramidal Side Effects Scale *Simpson, G. M., B., M., B., Gh., Angus, J. W. S., P., F. R. C., & M., D. P. (1970). A RATING SCALE FOR EXTRAPYRAMIDAL SIDE EFFECTS. Acta Psychiatrica Scandinavica, 45(S212), 11–19. https://doi.org/10.1111/j.1600-0447.1970.tb02066.x*						
Assess patients extrapyramidal side effects	10-item scale, each rated on a 5-point Likert scale. Total score is calculated by summing items, hence ranges from 0 – 40.	Varies for each item 0 being normal, 4 indicating abnormality in function being tested.	None found	Higher scores indicate more severe side effects i.e. "Higher worse"	None found	None found
International Personality Disorder Examination (IPDE)						
Designed to assess the personality disorders in the ICD-10 and DSM-IV classification systems.	77 item scale each item scored on a binary true or false. Total score for 10 personality disorders will be calculated by summing individual items on the scale. Please see Appendix 2 for more information	0 = false 1 = true	None found	n/a	None found	None found
Modified Overt-Aggression Scale (MOAS) *https://depts.washington.edu/dbpeds/Screening%20Tools/Modified-Overt-Aggression-Scale-MOAS.pdf*						
To rate the patient's aggressive behavior	4-item scale, each rated 0 – 4. For each subscale, frequency of 'event' is indicated, and multiplied by weight (item codes) for overall subscale score to give 'sum score'.	Varies for each item but ranges 0 - 4, 0 representing no aggression, 4 severe aggression.	• Verbal aggression (weight is 1) • Aggression against	Higher scores indicate more aggressive behaviour i.e. "Higher worse"	None found	None found

	'sum scores are then multiplied by weights (see subscales) and summed together for Total weighted score. Total score ranges from 0 with no upper limit, as frequencies of events indicate top score.	Item coding represents weight for the frequency of that event	property (weight is 2) • Auto aggression (weight is 3) • Physical aggression (weight is 4)			
Psychosis Screening Questionnaire						
Screening for psychosis	Measure not being scored or presented, screening questionnaire only	n/a	n/a	n/a	n/a	n/a
Standardised Assessment of Severity of Personality Disorder (SASPD)						
Provides a simple, brief, and reliable indicator of the presence of mild or moderate Personality Disorder according to ICD-11 criteria	9-item scale, each rated 0 – 3. Total score calculated by summing 9 items therefore, total score ranges from 0 to 27.	Each item is scored 0 = absent, 1 = mild, 2 = moderate, and 3 = severe	None found	Higher scores indicate more severe personality disorder i.e. "Higher worse"	None found	8 (mild PD) or 10 (moderate PD)
Structured Clinical Interview for Axis II Personality Disorders (SCID-II)						
To assess DSM-IV personality disorders	Measure not being scored or presented, screening questionnaire only	n/a	n/a	n/a	n/a	n/a
Zanarini rating scale for Borderline Personality Disorder (ZAN-BPD) *Zanarini rating scale for borderline personality disorder (ZAN-BPD): A continuous measure of DSM-IV Borderline Psychopathology Zanarini, Mary C Journal of Personality Disorders; Jun 2003; 17, 3; ProQuest pg. 233.* *Zanarini, M.C., & Frankenburg, F.R. (2001). Zanarini Rating Scale Borderline Personality Disorder (ZAN-PD).*						
Assessment of change in the DSM-IV BPD symptoms over time	9-item scale, each rated 0 – 4. Total score calculated by summing all items, therefore ranging from 0 – 36 *Subscales scored by summing relevant items for each.*	Items scored 0=no symptoms, 1 = mild symptoms, 2 = moderate symptoms, 3 = Serious symptoms and 4 = Severe symptoms	4 subscales: • Affective • cognitive • impulsive • interpersonal (disturbed relationship)	Higher scores indicate more severe symptoms i.e. "Higher worse"	None found	None found

APPENDIX 2 – IPDE SCORING SUMMARY

SCORING INSTRUCTIONS

1. *Circle the item numbers not followed by F, if were answered True.*
2. *Circle the remaining item numbers (those followed by F), if they were answered False.*
3. *If three or more items from a disorder are circled, the subject has failed the screen for the disorder, and should be interviewed.*
4. *Users of this instrument may wish to adopt lower or higher screening standards, depending on the nature of the sample, and the relative importance of errors in sensitivity (false negative cases) vs. specifically (false positive cases).*

301.0	Paranoid	2F	14F	36	38	58	66	72	
301.20	Schizoid	1F	12	21F	31	46	57F	77F	
301.22	Schizotypal	2F	24	30	52	64	67	70	71F	77F
301.50	Histrionic	5	10	17	26	28	35F	44	45	
301.7	Antisocial	11F	18F	20	29	47	56	74		
301.81	Narcissistic	7F	9	15	22	37	55	61	65	68
301.83	Borderline	4	8	13	25F	40	43	53	60	75
301.4	Compulsive	3F	19	23	32	41	48	54	59	
301.6	Dependent	6	33	42	49	50	62	69F	73	
301.82	Avoidant	16	27	34	38	39	51	63	76	