STATISTICAL ANALYSIS PLAN

FOR SUMMIT RCT

Date: 28/03/2025

Version: 1

Statistical analysis plan (SAP) for Implementing and evaluating group interpersonal therapy for postnatal depression in Lebanon and Kenya (**SUMMIT**: **SU**pporting **M**others' **M**ental health with **I**nterpersonal **T**herapy.

Funding Source:

NIHR's Research and Innovation for Global Health Transformation (RIGHT) programme

ISRCTN: ISRCTN52076264

UCL REC number: 18773/001

Funder ref: NIHR200851

This document has been written based on information contained in the study protocol version 1.1.

Chief Investigator: Professor Peter Fonagy

Local Chief Investigators:

- Lebanon: Dr Rabih Chammay
- Kenya: Dr Carol Ngunu

Trial Co-ordinators:

- Principal research coordinator Dr Liz Simes
- UK: Ciara O'Donnell & Sophie Wallace-Hanlon
- Lebanon: Sandra Pardi Maradian
- Kenya: Dr Beatrice Madeghe

Statistical co-applicant: Dr Zoë Hoare

Trial Statistician: Ms Rachel Evans

Rachel Evans 28/03/2025

Author's Name **Date and Signature**

Dr Zoë Hoare 28/03/2025

Second Statistical Advisor **Date and Signature**

Professor Peter Fonagy

Chief Investigator **Date and Signature**

28/03/2025

Document History

Updated version no.	Effective date	Authorship	Section changed	Summary of changes
1		R Evans	New	N/A

Acronyms and definition of terms

Acronym	Meaning
AE	Adverse Event
AR	Adverse Reaction
BISQ	Brief Infant Sleep Questionnaire
CI	Chief Investigator
CONSORT	Consolidated Standards for reporting Trials
CREDI	Caregiver Reported Early Development Index
CRF	Case Report Form
CTU	Clinical Trials Unit
CSI	Couple Satisfaction Index
DMEC	Data Monitoring and Ethics Committee
EDC	Electronic Data Capture
FCI	Family Care Indicator
g-IPT	Group Interpersonal Psychotherapy
GAD	Generalised Anxiety Disorder
GCP	Good Clinical Practice
HQ-SC	High Quality Standard Care
IPT	Interpersonal Psychotherapy
IQ	Installation Qualification
ITT	Intention to Treat
LMIC	low and middle-income countries
LSNS	Lubben Social Network Scale
MAR	Missing at Random
MCAR	Missing Completely at Random
MDAT	The Malawi Developmental Assessment Tool

MITT	Modified Intention to Treat
MMS	Modified Mini Screen
MNAR	Missing Not at Random
NWORTH	North Wales Organisation for Randomised Trials in Health
OQ	Operational Qualification
PHQ	Patient Health Questionnaire
PI	Principal Investigator
PID	Participant Identification
PND	Postnatal depression
PQ	Performance Qualification
QA	Quality Assurance
TMG	Trial Management Group
TSC	Trial Steering Committee
RCT	Randomised Controlled Trial
REDCap	Research Electronic Data Capture
SAE	Serious Adverse Event
SAP	Statistical Analysis Plan
SDV	Source Data Verification
SOP	Standard Operating Procedure
SUMMIT	SUpporting Mothers' Mental health with Interpersonal Therapy
UCL	University College London
URS	User Requirement Specification
VMP	Validation Master Plan

Table of Contents

1. Statistical analysis plan authorship

The analysis plan has been authored by Rachel Evans, Senior Statistician with input from Dr Zoë Hoare (statistical co-applicant), Dr Liz Simes (Central Trial Co-ordinator), Dr Liz Allison (London PI), Professor Peter Fonagy (CI), Dr Rabih Chammay (Lebanon PI/CI), Dr Carol Ngunu (Kenya CI), Dr Manasi Kumar (Kenya PI), Dr Andrew Nyandigisi (Kenya PI), Dr Lucina Koyio (Kenya PI), Dr Fouad Fouad (Lebanon PI), Professor Ghida Anani (Lebanon PI), Professor Pasco Fearon (London PI), Professor Stephen Pilling (London PI), Professor Henrietta Moore (London PI), Professor Jolene Skordis (London PI), Professor Lena Verdeli (Columbia University PI), Sandra Pardi Maradian (Lebanon Trial Co-ordinator), Dr Beatrice Madeghe (Kenya Trial Coordinator), Ciara O'Donnell and Sophie Wallace-Hanlon (London Trial Coordinators).

The draft plan will be circulated to the TSC and DMEC for comments before being agreed and signed off.

Statistical analysis will be completed by Rachel Evans at NWORTH with oversight from Zoë Hoare. Health economic analysis will be conducted by the Trial health economist Gerard Abou Jaoude. A separate analysis of qualitative data collected during the conceptual mapping will be conducted by Hannah Sender through the study. As agreed internally with Lebanon, Kenya & UCL teams, some of the qualitative data collected during the RCT will also be analyzed locally. Furthermore, adherence to the intervention (therapists and patients) will be analysed by partners in Columbia Therefore, this SAP details the analysis of quantitative measures excluding the health economic and adherence measures.

2. Introduction

2.1 Background and Rationale

Depression is the most common mental health issue affecting women of childbearing age. 20%-25% of women in low and middle-income countries (LMICs) experience depression during pregnancy or shortly after childbirth. This can be very distressing and affects not only the mother, but also her child. Women with depression often struggle to respond to their children's needs. Research shows that as a result of this children of women with postnatal depression (PND) have poorer learning, or cognitive development, and more emotional and behaviour problems as they grow up. This is especially true in LMICs, where families may also be struggling with many other challenges that can affect children's development negatively. Many women in LMICs have very little contact with healthcare services, so antenatal services can be a key opportunity to reach women in need of mental health support. However, currently treatment for PND is rarely available in many LMICs. The World Health Organisation recommends a therapy called interpersonal psychotherapy (IPT) to treat Depression (World Health Organization, 2016). There is research from high-income countries showing that IPT and group-IPT (g-IPT) is an effective treatment for PND but we do not know whether it works in a LMIC context, or whether it also benefits child development. This study aims to explore the effectiveness of g-IPT in two LMIC for women with PND.

2.2 Trial Aim

To assess whether or not culturally-adapted group interpersonal therapy (g-IPT) delivered in community settings in Kenya and Lebanon has a greater impact than high quality standard care (HQ-SC) on child developmental outcomes, maternal depression and the mother-child relationship.

2.3 Trial Population

Women with postnatal depression in Beirut, Lebanon and Nairobi, Kenya

Inclusion criteria:

- Aged 18 years or older
- Female
- Postnatal depression as indicated by a score of 12 or more on the PHQ-9 (Patient Health Questionnaire) at baseline (or during screening)
- Mother with an infant aged 6 - 35 weeks old at the time of screening

Exclusion criteria:

- Mothers with psychotic conditions including bipolar disorder, anorexia nervosa or substance dependency
- Mothers whose babies have severe physical health problems or neurodevelopmental problems will also be excluded

2.4 Trial Design

An individually randomised superiority trial of culturally adapted g-IPT versus HQ-SC for women with postnatal depression in Beirut, Lebanon and Nairobi, Kenya. Eligible mothers will be assessed at baseline and randomised to one of the two treatment conditions. Participants in both treatment conditions will be followed up for assessment by the research teams at 8, 13, 24, 36 and 52 weeks post first clinical contact. The primary outcome of the trial is The Malawi Developmental Assessment Tool (MDAT), see section 4.3 for a full list of outcomes.

Further information on the two treatment arms (g-IPT and HQ-SC) can be found in the trial protocol. The analysis will account for clustering in the intervention (g-IPT) group. Although HQ-SC is delivered in groups we will not need to account for clustering in these groups as the treatment occurs at baseline “pre-trial treatment” and across both arms of the trial.

3. Statistical Principles

Primary analysis will be on an intention-to-treat (ITT) basis, including all those randomised in the analysis set. The MDAT primary outcome measure will also be analysed on a per-protocol basis (see section 3.5 and 5.9), a modified intention to treat (MITT), excluding participants who only had baseline data. This is to assess any imputation assumptions made on participants that didn’t engage in the trial at all.

A complete case sensitivity analysis will also be run to assess impacts of multiple imputation, see section 5.9. Table 1 summarises the analysis sets and the applicable outcomes.

Table 1: Definitions of analysis sets

Analysis set	Definition	Outcome	Type
ITT	All those randomised analysed	Primary outcome and secondaries	Primary
MITT (1)	Excluding participants who only had baseline data	Primary outcome	Secondary sensitivity
MITT (2)	Excluding cases where index child has died	Primary outcome	Secondary sensitivity
Per-protocol	Excluding participants as defined in section 3.5	Primary outcome	Secondary sensitivity
Complete Case	All those complete data*	Primary outcome	Secondary sensitivity

*Complete data required for the analysis model i.e. outcome and all covariates.

3.1 Sample size justification

Power calculations indicate a sample of 412 (N = 224 in the g-IPT arm, and N = 188 to control) provides 90% power to detect a standardized mean difference of .40 on the primary outcome, taking into account up to 25% attrition. This assumes an effect size of 0.4 90% power, attrition of 25% and group size of 8 completing.

3.2 Randomisation

Participants will be randomised to receive control or intervention using a secure, web-based platform that can be accessed 24 hours a day, which was developed and maintained by NWORTH (Russell et al., 2011). 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. Allocation will be on a ratio (g-IPT: HQ-SC) of 1.33:1, stratified by site and age (18-21, 22-26, 27-31, 32>.)

3.3 Levels of confidence and p-values

All statistical tests and confidence intervals will be two-sided and performed using a 5% significance level and 95% confidence intervals will be presented.

3.4 Adherence

Adherence to the intervention (fidelity) by g-IPT therapists will be analysed by partners in Columbia. This is to explore fidelity to the g-IPT model. This includes data filled in by therapists (followed by randomization of sessions for each g-IPT group, where therapists were asked to fill in a detailed g-IPT supervision checklist based on the different phases (Pre-group meeting, initial phase, middle phase, termination phase).

Adherence to trial *treatment* during the trial (intervention and control) will be summarised descriptively in the quantitative analysis report, summarising number of sessions attended, length of the session and facilitator descriptives. In addition, sensitivity analysis will be conducted on this data, see section 5.9.

3.5 Protocol Violations and deviations

Violation is an intended failure to adhere to the protocol such as wrong treatment being administered, or incorrect data being collected and documented. A protocol deviation is an unintended failure to adhere to the protocol and examples include errors in applying inclusion/exclusion criteria or missed follow-up visits due to error.

Some examples of protocol deviations which might occur are;

- Participant becoming ineligible at a later date during the trial (i.e. diagnosis of neurological condition for the index child)
- Participant receives incorrect treatment (i.e. allocated to control and receives intervention)

A log of protocol deviations/violations will be kept by the trial co-ordinator and shared with the trial statistician at the end of the study to inform analysis data sets for any per protocol analysis to be conducted.

3.6 Missing Data

Levels of missing data will be monitored by the trial statistician throughout the data collection period. Where necessary the statistician will query missing data and where possible the missing data will be obtained by the trial co-ordinators. All methods to obtain missing data should be sought where possible. Completion rates of the outcome measures and other data will be calculated and presented in the final analysis report.

For missing items within a validated outcome measure, the published rules for completing missing data for the relevant measure will be applied, see Appendix 2. Where there are no missing data rules for the measure, if the number of missing items on an outcome 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 and multiple imputation methods will be used.

To investigate whether the data is missing completely at random (MCAR), Little (1998)’s missing completely at random test will be performed. To investigate whether the data is missing at random (MAR), explorative statistical tests (t-tests and chai square) will be conducted to assess if there any differences present between complete data and non-complete cases on specific variables indicating it is a predictor of missingness. Factors to be assessed as predictors of missingness include;

- Randomisation data (Site and Age group)
- Participant demographics (Age, religion, education, maternal socio-economic status, physical health questions, parity, single parent status and teen status)
- Family circumstances (marital status, number of adults and children living in the household and COVID-19 data)
- Initial Severity of depression symptoms (PHQ-9)
- Child birthweight

If data is indicated to be MCAR and/or MAR, then predictive mean matching multiple imputation method will be adopted. For multiple imputations, the number of imputations completed will be dependent upon the percentage of missing data (White et al., 2011). The missing outcome measures will be imputed using group allocation, stratification variables (i.e. Site and Age group) and any factors identified to be a predictor of missingness.

If the data is evaluated to be MNAR then additional modelling guided by clinical knowledge would be required to simulate the missing data mechanism and impute the missing data guided by any indicated variables as systemically different. Any methods used will be clearly detailed in the final report. Primary analysis will be conducted on the imputed dataset and sensitivity analysis will be conducted on the complete case data if required. Sensitivity analysis will also be conducted on an MITT basis, excluding cases where the index child has died as MI may not be appropriate in these instances.

This is only applicable for the primary outcome, MDAT. Other secondary outcomes will be run with linear mixed models across timepoints, and multiple imputation is not required. See section 5.6 and 5.7 for more detail.

3.7 Assumption Checking

All assumptions relating to the models will be checked and evaluated whether appropriate to use with the data. If any of the assumptions are substantially violated, then appropriate non-parametric tests will be conducted. Table 2 contains details of the assumptions associated with each model and the methods to be used to assess these assumptions.

Table 2: Assumptions of analysis models to be checked

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
Residuals/Errors are normally distributed	P-P plot/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

During monthly data cleaning, any outliers identified will be queried by the Trial Statistician with the Trial co-ordinators. This will be to identify whether the outlier is a data entry error, a measurement error or to confirm that it is a genuinely unusual value. Once this has been clarified the data will be amended if necessary or will remain unchanged if identified to be correct. No outliers will be discarded from analysis if they are within range.

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

4. Data

Data collection and entry onto a REDCap (Harris et. al., 2019) database will be undertaken at sites (Lebanon and Kenya). Cleaning and analysis will be undertaken by NWORTH using standard, secure, anonymous procedures for handling research data supported by the central research team at UCL. The fully auditable REDCap data management system will be used to ensure best practice. For full details on the data collection, flow and storage please refer to the current version of the SUMMIT Data Management Plan.

4.1 Data Collection and handling

Quantitative research data will be collected via laptops or tablets, entered directly onto the REDCap database. There will be back-up paper CRFs in case there is a problem with either the REDCap database or the tablet being used. If the data is collected on paper CRFs these will be entered onto the REDCap database by the research assistant as soon as possible after the data has been collected. Paper versions of the data will be kept in locked filing cabinets, separate from any identifiable data such as consent forms, at local sites in accordance with Good Clinical Practice (GCP).

Researchers (data collectors and trial co-ordinators) will be collecting the data, primarily via face-to-face interviews or via telephone interviews if face-to-face is not possible.

4.2 Time points of outcome measures

Table 3 contains the full list of study outcome measures and their time point collections.

Table 3: Outcome measures collection for SUMMIT according to Trial Protocol

Outcomes measure	**Screening (Pre-treatment assessment)**	**Baseline (T1)**	**8 weeks (T2)**	**13 weeks (T3)**	**24 weeks (T4)**	**36 weeks (T5)**	**52 weeks (T6)**
1. Patient Health Questionnaire – depression module (PHQ-9)	X	X*	X	X	X	X	X
2. Whooley questions	X**						
3. The Modified Mini Screen (MMS) – section C only	X						
4. Generalised Anxiety Disorder Assessment (GAD7)		X	X	X	X	X	X
5. The Malawi Developmental Assessment Tool (MDAT)							X
6. EQ-5D-5L		X		X	X		X
7. ICECAP- A questionnaire		X		X	X		X
8. SUMMIT patient careseeking and costs				X			
9. The Caregiver Reported Early Development Index (CREDI) long form		X		X		X	
10. Brief Infant Sleep Questionnaire – Revised Short form (BISQ)		X	X	X	X	X	X
11. Infant physical health questionnaire		X	X	X	X	X	X
12. Breastfeeding outcome measure		X	X	X	X	X	X
13. Sleep Condition Indicator		X	X	X	X	X	X
14. The Lubben Social Network Scale (LSNS-6)		X		X			X
15. Couple satisfaction Index (CSI-4)		X		X			X
16. Demographic questionnaire: • Socio-economic status • Maternal education • Maternal parity		X					

Outcomes measure	Screening (Pre-treatment assessment)	Baseline (T1)	8 weeks (T2)	13 weeks (T3)	24 weeks (T4)	36 weeks (T5)	52 weeks (T6)
Teen parent status							
17. Family circumstances questionnaire		X		X			X
18. Semi-structured interview exploring the mothers' experience of postnatal depression		X		X			X
19. Family Care Indicator (FCI)		X	X	X	X	X	X
20. Household economic questionnaire		X					
21. Household shocks questionnaire				X			

*The PHQ-9 will only be collected at baseline if it has been 7 or more days since the screening PHQ-9 was completed.

**The Whooley questions will only be completed in Kenya.

4.3 Definitions and calculations of outcome measures

For information on the scoring of some of the outcome measures see Appendix 2.

4.4 Safety Data

For definitions and details of safety reporting please refer to the study Protocol. Any safety events will be reported with details of severity, cause and outcome included. This will be presented overall and by group. The number and percentage of patients that have been affected by an adverse event will also be presented. No formal statistical testing will be undertaken on these data.

5. Statistical analyses

5.1 Analysis Time Frame

Table 4: Summary of expected analysis timelines

TASK	EXPECTED DATE
First participant randomised	January 2023
Final participant randomised	October 2023
Final participant followed up	January 2025
Data locking*	March 2025
Analysis completed**	April - June 2025

**The completion of data lock is dependent on the date of the final follow up, data entry from all sites and responses to data queries from sites.*
***a minimum of 1 month between data lock and analysis report delivery is required.*

5.2 CONSORT Analysis

The patient flow information, as advised by CONSORT reporting standards (Schulz et al., 2010) will be completed with participants numbers, as shown in Figure 1. Eligibility rates, recruitment rates, and retention rates will be reported using this data. Furthermore, details on reasons for ineligibility, non-consent and non-randomisation will be reported within a table along with their related patient frequencies and percentages. Reasons for withdrawal and lost-to-follow up will be presented where applicable and the associated time point of withdrawal or loss during the trial.

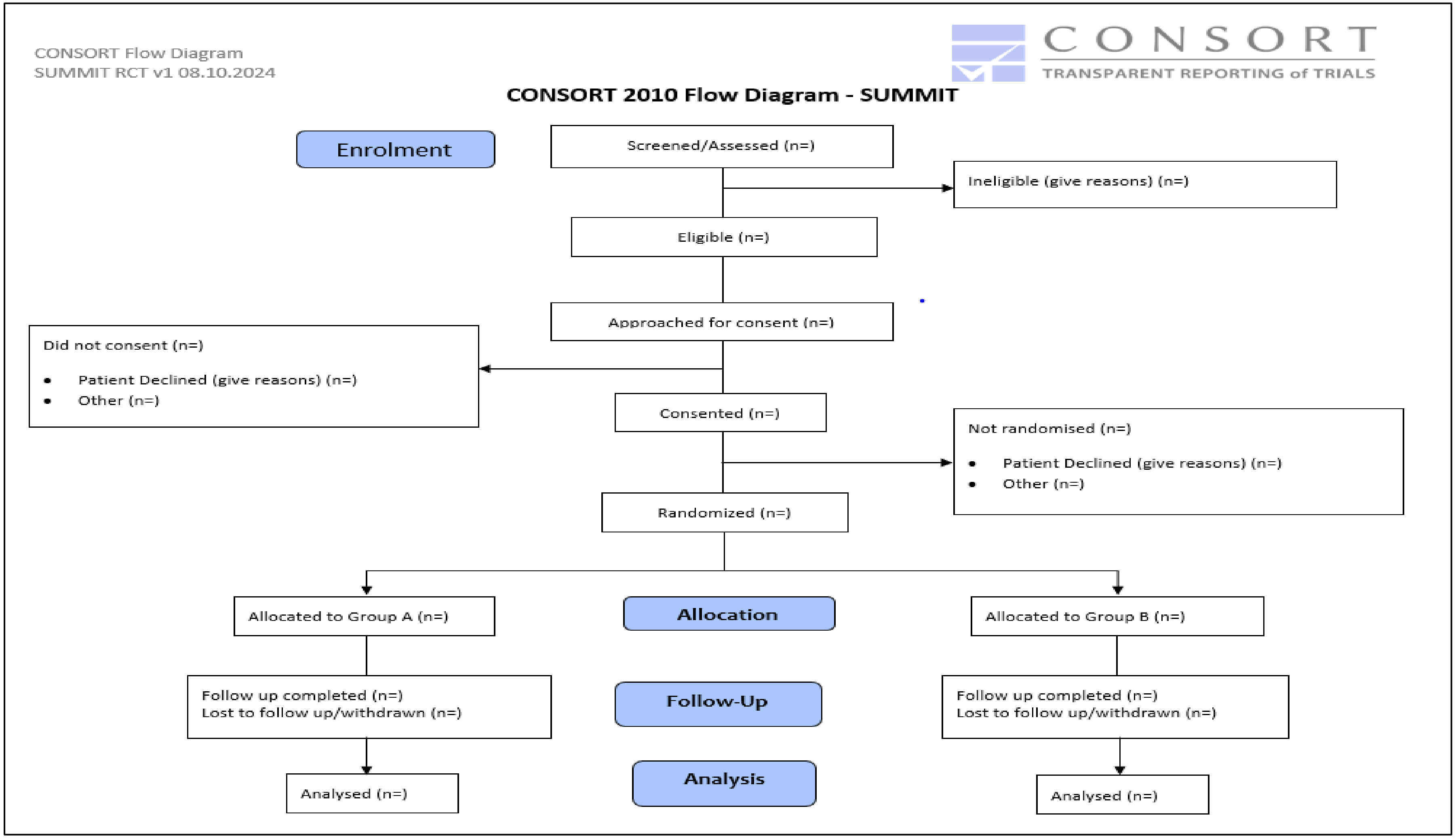

CONSORT Flow Diagram
SUMMIT RCT v1 08.10.2024
CONSORT
TRANSPARENT REPORTING of TRIALS
CONSORT 2010 Flow Diagram - SUMMIT
Enrolment
Screened/Assessed (n=)
Ineligible (give reasons) (n=)
Eligible (n=)
Approached for consent (n=)
Did not consent (n=)
Patient Declined (give reasons) (n=)
Other (n=)
Consented (n=)
Not randomised (n=)
Patient Declined (give reasons) (n=)
Other (n=)
Randomized (n=)
Allocated to Group A (n=)
Allocation
Allocated to Group B (n=)
Follow up completed (n=)
Lost to follow up/withdrawn (n=)
Follow-Up
Follow up completed (n=)
Lost to follow up/withdrawn (n=)
Analysed (n=)
Analysis
Analysed (n=)

Figure 1: Patient flow diagram for SUMMIT Trial, guided by CONSORT guidelines.

5.3 Baseline Analysis

A separate baseline analysis will not be conducted. A section of the main analysis and report will detail the characteristics of the study sample at baseline. No formal statistical testing will be conducted at baseline as indicated by CONSORT (Begg et. Al 1996) statistical testing at baseline is not informative. Descriptive statistics will be used to describe any imbalance, see section 5.5.

5.4 Interim Analysis

There is no planned interim analysis for the study.

5.5 Descriptive Statistics

Descriptive statistics of the data will be presented in the final analysis report. This will include randomisation figures, demographics and descriptive statistics of the outcome measures. All will be presented overall and split by Country and treatment allocation (group).

The following demographic descriptive statistics and other study variables will be presented;

- Randomisation data (Site and Age group)
- Participant demographics (Age, religion, education, maternal socio-economic status, physical health questions, parity, single or teen status and other indicators of low social support)
- Family circumstances (marital status, number of adults and children living in the household and COVID-19 data)

Descriptive statistics will be produced for the primary and all other secondary outcome variables and data at respective timepoints, as listed in Table 3, section 4.2.

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

5.6 Analysis of Primary Outcome

The primary outcome, the MDAT, will be analysed using a general linear mixed model to assess the differences between allocation groups (g-IPT and HS-QC) at 52 weeks. The primary analysis will account for the clustering in the intervention arm by including g-IPT group in the model. The stratification variables (Site and Age group) along with Severity of Depression (PHQ-9), and index child's birth weight will also be included in the model.

Multiple imputation techniques described in section 3.6 will be utilised on the MDAT for the primary analysis in line with an ITT population. A complete case sensitivity analysis will also be run, see section 5.9. Further, a modified ITT analysis will be carried out on the MDAT, excluding participants who only had baseline data. This is to assess any imputation assumptions made on participants that didn't engage in the trial at all. Primary analysis will be on an intention-to-treat basis, subsequent per protocol analysis will also be run if required (see section 3.5).

5.7 Analysis of Secondary Outcomes

For all secondary outcomes analysis will be by Linear mixed modelling taking account of clustering by g-IPT group in the treatment arm, controlling for stratification factors (site and age group) and covariates Severity of Depression (PHQ-9), and index child's birth weight, and modelling longitudinal effects of time, and time of treatment interactions where appropriate. The outcomes will be analysed with an analogous model appropriate for the outcome type including baseline measures where appropriate in addition to the stratification and pre-defined covariates.

All results will be presented without an adjustment for multiple comparisons however when any conclusions drawn from results multiple comparison effects will be taken into account.

Appendix 1 details all outcomes for the trial, variable types and their corresponding analysis.

5.8 Subgroup Analysis

Subgroup analysis on the primary outcome (MDAT at 24 weeks, ITT analysis set) will be run for the Country subgroups (Lebanon/Kenya).

Any covariates found to be consistently important in the main effects models will be analysed for subgroup effects.

Additionally, the following variables will be explored for subgroup analysis:

- maternal education,
- maternal parity,
- single parent status,

5.9 Sensitivity Analysis and model testing

Sensitivity analysis will be conducted on the primary outcome, the MDAT at 24 weeks for the below:

- Analysis sets (mITT, complete case and per protocol as described in Section 3 Table 1)
- Adherence to treatment, variables indicating number of sessions participant attended will be included in the primary model.*
- The possible impact of major geopolitical events, likely to affect recruitment, treatment delivery or outcomes will be analysed in sensitivity analysis.
- Analysis only including the participants whose data was collected within the aimed timeframe.
- Economic status (as defined below) will be included in the analysis model to evaluate impacts on the primary outcome.

The health economist will estimate the socio-economic status of SUMMIT participants, which will be used as a covariate for sensitivity analysis. A composite index of socio-economic status will be generated based on cross-sectional baseline survey data on participant and household characteristics, including education, overcrowding, assets and *income. Principle component analysis (PCA) will be employed, separately for Kenya and Lebanon, to select variables that will form part of a composite socio-economic index for each country. SUMMIT participants will then be disaggregated into socio-economic groups (e.g. quintiles) based on their estimated socio-economic index. The generation of a socio-economic index is contingent on the completeness of baseline data. If data are missing, in most cases (e.g. educational status) it is unlikely that sufficient information will be available to impute missing data. Such instances will likely require a variable to be excluded from the PCA. In the event that substantial data are missing across a number of baseline variables, another approach will be employed to estimate socio-economic status using a single variable (e.g. income) or a single type of variables (e.g. asset-based index).*

Sensitivity analysis was considered for where researchers collecting outcome data become unblind. This only occurred in a few cases in Lebanon and none in Kenya. Where researchers were unblinded another researcher conducted the follow ups therefore sensitivity analysis is not required.

5.10 Exploratory analysis

Mediation of mood, social factors, and on primary outcome.

A causal mediation analysis (including mood, social factors, and adherence to treatment (fidelity) as mediators in separate models) will be implemented using a counterfactual framework to include a treatment/mediator interaction and covariates, based on univariable path screening, and with bias corrected bootstrapped estimates.

Mediator measures;

- Mood (PHQ-9 and GAD7)

- Social factors (LSNS-6 and FCI)
- Adherence to treatment protocol (fidelity)

Moderator measures**;**

- Severity of Depression (PHQ-9)
- index child's birth weight

5.11 Unblinding

Due to the unequal allocation the Trial Statistician conducting analysis will be unblind to the treatment group. The statistical analysis report will result in unblinding of blinded team members therefore when the statistician shares the results with team members (via report or results presentation) this will be considered the point at which those members are unblind for the Trial.

6. Software

All quantitative analysis will be completed using Stata 18, SPSS v25 and R version 3 or higher. Analysis code can be made available to the CI when issuing the data pack at the end of the study. Analysis code for the models defined in this SAP will not require verification by a second statistician. If subsequent code or exploratory analysis is more complex, then this analysis code would be verified by a second statistician and documented as outlined by NWORTH procedures (5.WI.03).

7. References

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., ... & Stroup, D. F. (1996). Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *Jama*, *276*(8), 637-639.

Bono C, Ried LD, Kimberlin C, Vogel B. Missing data on the Center for Epidemiologic Studies Depression Scale: a comparison of 4 imputation techniques. Res Social Adm Pharm. 2007 Mar;3(1):1-27. doi: 10.1016/j.sapharm.2006.04.001. PMID: 17350555.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics, 11*(1), 1-21.

PA Harris, R Taylor, BL Minor, V Elliott, M Fernandez, L O'Neal, L McLeod, G Delacqua, F Delacqua, J Kirby, SN Duda, REDCap Consortium, The REDCap consortium: Building an international community of software partners, J Biomed Inform. 2019 May 9 [doi: 10.1016/j.jbi.2019.103208]

Russell, D., Hoare, Z., Whitaker, R., Whitaker, C., & Russell, I. (2011). Generalized method for adaptive randomization in clinical trials. Statistics in medicine, 30(9), 922-934.

Schulz KF, Altman DG, Moher D, for the CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. BMC Medicine 2010, 8:18. (24 March 2010)

StataCorp. 2021. Stata Statistical Software: Release 17. College Station, TX: StataCorp LLC.

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.

World Health Organization. (2016). Group interpersonal therapy (IPT) for depression.

Referenced documents:

Referenced documents:

1. SUMMIT Data Management Plan v1 20/03/2023

2. SUMMIT RCT protocol v1.1

8.Appendices

Appendix 1 – Study outcomes and analysis type

Outcome	No. of variables	Variable type	Timepoint(s) for analysis	Method
1. Patient Health Questionnaire – depression module (PHQ-9)	1 total score	Continuous	All as collected, see table 3	Across time
2. Whooley questions	N/A	N/A - Screening measure only – not being analysed	N/A - Screening measure only – not being analysed	N/A - Screening measure only – not being analysed
3. The Modified Mini Screen (MMS) – section C only	N/A	N/A - Screening measure only – not being analysed	N/A - Screening measure only – not being analysed	N/A - Screening measure only – not being analysed
4. Generalised Anxiety Disorder Assessment (GAD7)	1 total score	Continuous	All as collected, see table 3	Across time
5. The Malawi Developmental Assessment Tool (MDAT)	1 total score	Continuous	52-week endpoint	At endpoint
6. EQ-5D	N/A	N/A - Health economic	N/A - Health economic	Health economics measure – not analysed by NWORTH
7. ICECAP- A questionnaire	N/A	N/A - Health economic	N/A - Health economic	Health economics measure – not analysed by NWORTH
8. SUMMIT patient careseeking and costs	N/A	N/A - Health economic	N/A - Health economic	Health economics measure – not analysed by NWORTH
9. The Caregiver Reported Early Development Index (CREDI) long form	1 total score	Continuous	All as collected, see table 3	Across time
10. Brief Infant Sleep Questionnaire – Revised Short form (BISQ)	2 subscale scores	Continuous	All as collected, see table 3	Across time
11. Infant physical health questionnaire	N/A	N/A - Descriptive only	N/A - Descriptive only	N/A - Descriptive only
12. Breastfeeding outcome measure	N/A	N/A - Descriptive only	N/A - Descriptive only	N/A - Descriptive only
13. Sleep Condition Indicator	1 total score	Continuous	All as collected, see table 3	Across time
14. The Lubben Social Network Scale (LSNS-6)	1 total score	Continuous	All as collected, see table 3	Across time
15. Couple satisfaction Index (CSI-4)	1 total score	Continuous	All as collected, see table 3	Across time
16. Demographic questionnaire:	N/A	N/A - Descriptive only	N/A - Descriptive only	N/A - Descriptive only
17. Family circumstances questionnaire	N/A	N/A - Descriptive only	N/A - Descriptive only	N/A - Descriptive only
18. Semi-structured interview exploring the mothers' experience of postnatal depression	N/A	N/A - Qualitative	N/A - Qualitative	Qualitative data – not being analysed here
19. Family Care Indicator (FCI)	6 subscale scores	Continuous	All as collected, see table 3	Across time

Outcome	No. of variables	Variable type	Timepoint(s) for analysis	Method
20. Household economic questionnaire	N/A	N/A - Health economic	N/A - Health economic	Health economics measure – not analysed by NWORTH
21. Household shocks questionnaire	N/A	N/A - Health economic	N/A - Health economic	Health economics measure – not analysed by NWORTH

*No. of variables refers to number of scores generated from the measure for analysis in line with Appendix 2

Appendix 2 – Scoring of Validated outcome measures

Definition	Item Coding	Scoring	Subscales	Direction	Missing value rules	Thresholds
Patient Health Questionnaire – depression module (PHQ-9) *Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: A New Depression Diagnostic and Severity Measure. Psychiatric Annals, 32(9), 509–515.*						
depression rating scale	9 item scale. Items scored on a 4-point Likert scale from 0 to 3. **(Note: items in database are coded 1 to 4, need re-coding 0 to 3 before scoring)**	Item scores are summed together to calculate the total score ranging from 0 to 27.	None identified	Higher scores indicates more depressive symptoms i.e. higher scores worse	None found	<10 no major depression >15 major depression 10 to 14 grey zone
Generalised Anxiety Disorder Assessment (GAD7) *Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7. Archives of Internal Medicine, 166(10), 1092. https://doi.org/10.1001/archinte.166.10.1092*						
assessing generalized anxiety disorder	7 item scale. Items scored on a 4-point Likert scale from 0 to 3.	Item scores are summed together to calculate the total score ranging from 0 to 21.	None identified	Higher scores indicate higher anxiety severity i.e. higher scores worse	None found	0–4: minimal 5–9: mild 10–14: moderate 15–21: severe
The Malawi Developmental Assessment Tool (MDAT) *The MDAT scoring information was requested from the developers*						
	136 item scale. Items scored either 0 (Fail) or 1 (Pass) Pass (1) = Yes (1) Fail (0) = No (0)	An App to score the measure has been created by developers. https://kieran-bromley.shinyapps.io/mdat_scoring_shiny/ Z scores can be downloaded from the app for analysis and interpretation	Gross Motor (GM) (34 items) Fine Motor and Performance (FM) (34 items) Language/Hearing (34 items) Social (34 items)	Higher scores indicate better response i.e. more "passes"	Scoring app handles missing data. If completely missing, then they have no score generated. If a child has some missing items, score generated as	None found

		The model being used is a two-parameter-logistic model to estimate the latent construct (development) given the item responses (to the MDAT). This model focuses on the probability of a subject passing an item given their level of development, and where there is a missing response, it does not contribute to the likelihood.			their response vector can still be used to estimate their development score.	
The Caregiver Reported Early Development Index (CREDI) long form *https://credi.gse.harvard.edu/files/credi/files/credi_scoring_manual_15-october-2021.pdf* *CREDI Long Form produces an overall developmental score, as well as scores for each developmental domain: motor, cognitive, language, and social-emotional. (A scoring system for mental health is pending.) The questionnaire data can be scored using either the credi package in R or the CREDI Scoring App.* https://inee.org/sites/default/files/resources/CREDI_Scoring_Manual_Eng.pdf						
Steps to score are listed below as indicated in the above referenced manual. 1. Name CREDI variables correctly 2. Ensure variables are coded properly 3. Remove personally identifiable information from your data 4. Save your data as an .xlsx or .csv file 5. Access the CREDI Scoring App 6. Indicate whether your data are already reverse-coded or not 7. Indicate if you wish to preserve item-level responses 8. Upload and score data 9. Download data 10. Examine scoring outcomes Further information: the scoring app does not consider the stopping rule when calculating scores. The main goal of the stop rule was to reduce administration time. In theory, continuing beyond the stop rules should not affect the generated scores much, but will actually (slightly) reduce the error associated with any one observation. This is because we are using a Bayseian multi-dimensional item factor analysis model to generate domain-level scores (described here) and a simpler 2PL IRT model to generate "Overall" scores (described here). In principle, these methods should assure that the addition of more items does not bias scores. Because items administered *after* 5 No/DK responses are likely to be *very* difficult for the child, it's likely that they are providing very little information about the child's developmental level.						

The Mental Health items do not have great psychometric properties according to our analyses and the app does not officially endorse any specific way of calculating scores. Missing responses are treated as missing in the scoring algorithm. In the background, the scoring app calculated expected-a-posteriori (EAP) scores given 1) the age of the child and 2) observed item responses and 3) item parameters. The generated score is thus the median of the posterior distribution, and the standard error of measurement is the standard deviation of the posterior distribution.						
Sleep Condition Indicator *Espie, C. A., Kyle, S. D., Hames, P., Gardani, M., Fleming, L., & Cape, J. (2014). The Sleep Condition Indicator: A clinical screening tool to evaluate insomnia disorder: BMJ Open, 4(3), e004183. https://doi.org/10.1136/bmjopen-2013-004183*						
evaluate insomnia disorder	8 item scale. Each item scored on a 5-point Likert scale, reversed scored from 4 to 0.	Total score calculated by summing items therefore ranges from 0 to 32. Scores can be converted to 0-10 format (minimum 0, maximum 10) by dividing total by 3.2 to facilitate interpretation	None identified	Higher score means better sleep i.e. higher scores better	None found	Individual item score of 0, 1 or 2 represents Insomnia Disorder
The Lubben Social Network Scale (LSNS-6) Lubben, J., Gironda, M. (2004). Measuring social networks and assessing their benefits. In Social Networks and Social Exclusion: Sociological and Policy Perspectives. Eds. Phillipson, C., Allan, G., Morgan, D. Ashgate. Scoring obtained from: *https://www.brandeis.edu/roybal/docs/LSNS_website_PDF.pdf*						
self-report measure of social engagement including family and friends	6 item scale. Each item scored from 0 to 5.	Total score is calculated by summing the 6 items therefore the total score ranges from 0 to 30.	None found	higher score indicates more social engagement i.e. higher scores better	None found	None found
Couple satisfaction Index (CSI-4) Rogge, Ronald. (2007). The Couples Satisfaction Index: CSI-4. 10.13140/RG.2.1.4198.3129. https://www.researchgate.net/publication/299432417_The_Couples_Satisfaction_Index_CSI-4/link/56f68d0508ae38d710a1bbd7/download						

Measure of relationship satisfaction	4 item scale. Item 1 scored on a 6-point Likert scale Items 2-4 scored on a 5-point Likert scale with some reverse scored.	Total score is calculated by summing the 4 items therefore the total score ranges from 0 to 21	None found	higher scores indicate more satisfaction i.e. higher scores better	None found	scores falling below 13.5 suggest notable relationship dissatisfaction
Family Care indicator (FCI)* Kariger, P., Frongillo, E, A., Engle, P., Britto, R., Sywulka, S, M. & Menon, P. (2012). Indicators of family care for development for use in multicountry surveys. Journal of Health, Population and Nutrition, 30(4), 472-486.						
Measure of family care	variety	Five subscales to be calculated with additional 3 items on harsh parenting (not to be combined) Each subscale is calculated by summing items. See Table 5 for further scoring information. A total score for the measure will not be calculated	varieties of play materials (7 items) which classified toys by their use	higher better	None found	None found
			Sources of play materials (4 items) which identified where the play materials came from	higher better		
			play activities (6 items) which identified specific types of activities done by any adult in the home with the child in the previous three days	higher better		
			Household books (1 item) the number of books in the home, excluding picture books for young children household books	higher better		
			Magazines (1 item) i.e. the number of magazines	higher better		

			and newspapers in the home			
			harsh parenting practices (3 items)	Higher worse		
Brief Infant Sleep Questionnaire – Revised Short form (BISQ-R SF) https://www.babysleep.com/wp-content/uploads/2020/06/BISQR-agreement-5_26_2020-1.pdf						
Measure of infant sleep	variety	Two subscale scores will be calculated for the measure. These are both continuous measures. A Total Score will not be calculated, other data collected in the measure will be used descriptively.	• Nocturnal sleep (hours of sleep per night), Chspendsleephrs & chspendsleepmin • number of nighttime wakings chwakengt	Higher scores better Higher scores worse	None found	None found

**Not all items are available for the FCI as the measure was adapted for cultural relevance, see below for specific details on FCI scoring.*

Health Economics measures

Some measures are health economics measures and will not be scored at NWORTH. Raw item data will be included in the health economics dataset and end of study data pack. This is for the EQ-5D-5L, ICECAP-A questionnaire, Household questionnaire –shocks, and SUMMIT patient careseeking and costs.

Non scored measures:

The MMS and Whooley measures are screening measures only and will not be entered into the study database as they do not require analysis.

Infant physical health questionnaire, Breastfeeding outcome measure, Demographics and the Family circumstances questionnaire are all measures that do not require scoring. This data will be summarised descriptively where appropriate in the results report. Furthermore, many items were collected for the BISQ measure and only 3 are being used for outcome analysis therefore the rest of the data collected will be presented descriptively. Finally, the Semi-structured interview will be used for the qualitative analysis and is not part of the current statistical analysis.

FCI measure

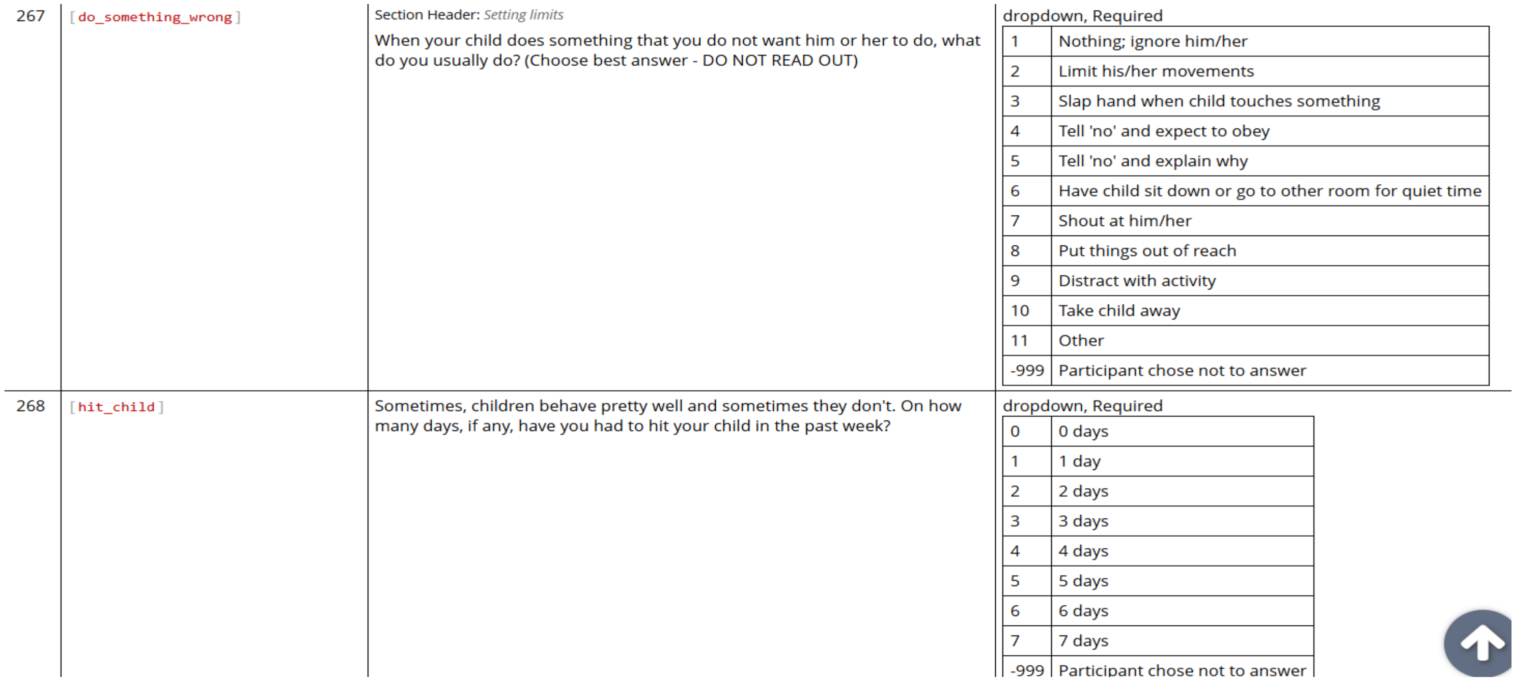
As the FCI measure was adapted for cultural relevance in the trial, there are a few items from the validated scoring that are not available in the data. Table 5 summarises how the measure is to be scored in line with the validated measure referenced above and indicated which items are available for the SUMMIT trial. Where we haven't got the data, the item will be ignored and the scale summed with the available items. The measures will be referred as “modified” in any publications and results write up and it will be clear that we scored them with some items in the scales missing.

Table 5: Specific scoring information for the FCI

Scale	Scale item	Item in data?	REDcap variable	Item level coding	Total range of scale
Varieties of play materials	Things which make/play music	yes	250	0 No, 1 yes	1 - 11
	Things for drawing/writing	yes	252	0 No, 1 yes	
	Picture books for children (not school-books)	yes	245	1 –6	
	Things meant for stacking, con structing, building (blocks)	yes	251	0 No, 1 yes	
	Things for moving around (balls, bats, etc.)	yes	253	0 No, 1 yes	
	Toys for learning shapes and colours	no	X	n/a	
	Things for pretending (dolls, tea-set, etc.)	yes	254	0 No, 1 yes	
Play activities	Read books or look at picture-books with child	yes	256	0 – 7	0 - 42
	Tell stories to child	yes	257	0 – 7	
	Sing songs with child	yes	258	0 - 7	
	Take child outside home place	yes	259	0 - 7	
	Play with the child with toys	yes	260	0 - 7	
	Spend time with child in naming things, counting, drawing	yes	261	0 - 7	
Sources of play materials	Household objects	yes	248	0 No, 1 yes	0 - 3
	Things from outside	yes	249	0 No, 1 yes	
	Toys bought from store	no	X	n/a	
	Home-made toys	yes	247	0 No, 1 yes	
Household books		yes	246	1-6	1 - 6

Magazines		no	X	n/a	n/a

In addition to the scales in Table 5 a “harsh parenting scale” will be created which will be yes/no binary variable. Participants will be classified as “yes” on this scale if the response to question 267 is 2, 3 or 7 and/or if the response to question 268 is > than 0 (i.e. 1 – 7), see figure below for information on questions 27 and 268.



267 [do_something_wrong]

Section Header: *Setting limits*

When your child does something that you do not want him or her to do, what do you usually do? (Choose best answer - DO NOT READ OUT)

dropdown, Required

1	Nothing; ignore him/her
2	Limit his/her movements
3	Slap hand when child touches something
4	Tell 'no' and expect to obey
5	Tell 'no' and explain why
6	Have child sit down or go to other room for quiet time
7	Shout at him/her
8	Put things out of reach
9	Distract with activity
10	Take child away
11	Other
-999	Participant chose not to answer

268 [hit_child]

Sometimes, children behave pretty well and sometimes they don't. On how many days, if any, have you had to hit your child in the past week?

dropdown, Required

0	0 days
1	1 day
2	2 days
3	3 days
4	4 days
5	5 days
6	6 days
7	7 days
-999	Participant chose not to answer