Statistical Analysis Plan (SAP):
Comparing the Effectiveness of Clinicians and Paraprofessionals to Reduce Disparities in Perinatal Depression

Version 5.0
August 6, 2019

Principal Investigator:
S. Darius Tandon, PhD
Associate Professor
Department of Medical Social Sciences
Northwestern University
750 N Lake Shore Drive, Suite 675
Chicago, IL 60611
312.503.3398
dtandon@northwestern.edu

Statistical Team:
Jody D. Ciolino, PhD
Assistant Professor
Division of Biostatistics/Biostatistics Collaboration Center
Department of Preventive Medicine
Feinberg School of Medicine
Northwestern University
680 N lake Shore Drive, Suite 1400
Chicago, IL 60611
jody.ciolino@northwestern.edu

Chen Yeh, MS
Statistical Analyst
Division of Biostatistics/Biostatistics Collaboration Center
Department of Preventive Medicine
Feinberg School of Medicine
Northwestern University
680 N lake Shore Drive, Suite 1400
Chicago, IL 60611
jyeh0909@northwestern.edu

Introduction

This document outlines the proposed analyses for the PCORI-funded "Mothers and Babies" study: a cluster-randomized study comparing the effectiveness of clinicians and paraprofessionals to reduce disparities in perinatal depression. Previously, the investigators of this study have established efficacy of the home visiting (HV) program: Mothers and Babies (MB) Course in preventing onset of postpartum depression and reducing depressive symptoms when led by mental health (MH) professionals. However, to date there are no interventions led by non-health or non-mental health professionals that have demonstrated efficacy in preventing the onset and worsening of postpartum depression among low-income women. This project attempts to fill this notable gap. We will conduct a cluster-randomized trial in which HV clients receive either a) MB delivered by mental health professionals (MH), b) MB delivered by paraprofessional home visitors (HV), or c) usual home visiting services. This study design will allow us to conduct a superiority trial that compares the efficacy of MB delivered by paraprofessional home visitors versus usual care. A superiority trial will allow us to generate efficacy data on MB delivered by paraprofessional home visitors. Our study design will also allow us to conduct a non-inferiority analysis that compares the effectiveness of MB delivered by mental health professionals versus paraprofessional home visitors. Should we find that paraprofessional home visitors are not inferior to mental health professionals in delivering the intervention, HV programs throughout the United States may be able to implement the MB Course with paraprofessional home visitors-an approach that is considerably more efficient and cost-effective than employing mental health professionals.

The study will employ a modified covariate-constrained cluster-randomized design at the site level, using unequal (1:3:3; control: mental health professional delivery of MB: paraprofessional delivery of MB) allocation, with intention to achieve relative balance in a set of pre-specified potential covariates of interest.

Specific Aim #1 is to conduct a superiority trial that compares the efficacy of MB delivered by paraprofessional home visitors versus usual care (i.e., home visiting without MB) on patient-reported outcomes, including depressive symptoms, quality of life, parenting practices, engagement in pleasant activities, and relationship with one's partner.

Specific Aim #2 is to conduct a non-inferiority trial that compares the effectiveness of MB delivered by (a) mental health clinicians versus (b) paraprofessional home visitors.

Specific Aim #3 is to evaluate whether effectiveness of the two versions of MB (clinician led vs. paraprofessional home visitor led) varies according to patient characteristics (e.g., race, ethnicity, first-time mother, and/or geographic type of home visiting (HV) program (i.e., urban vs. rural).

Specific Aim #4 is to examine the feasibility and acceptability of MB delivered by paraprofessional home visitors and mental health clinicians.

Study time points include: baseline, immediately following the intervention period (or roughly eight weeks post enrollment for controls), 12 weeks postpartum, and 24 weeks postpartum. Primary time point of interest will be 24 weeks postpartum. This analysis plan will focus on the primary results manuscript, and thus, we will focus on Aims #1-#3.

Study Outcomes

Primary Outcome

Primary outcome of interest for Aims #1-#3 will be the Quick Inventory or Depressive Symptoms (QIDS-SR16) score as determined by participant self-report at 24 weeks postpartum. We will control for baseline QIDS-SR16 score. This is a continuous measure ranging from zero to 27 points, where higher scores signify increased depressive symptoms. QIDS-SR16 translates into depressive categories such that a score of less than five points indicates no depression, a score ranging from six to 10 indicates mild depression, 11-15 signifies moderate depression, 16-20

indicates severe depression, and anything above 20 would be labeled as very severe (http://www.ids-qids.org/). As a result, we deem a five-unit change or difference in score to be meaningful (as a jump in five points would result in an increase in depression severity tier for any individual patient).

Secondary Outcomes

The following secondary outcomes will further be examined in order to address subcomponents of specific aims and also to address some of the exploratory aims. These secondary outcomes align with key components of the Mothers and Babies intervention.

1. Incidence of major/clinical depression (major depressive episode [MDE])
    a. A major depressive episode will be defined via the Maternal Mood Screener
    b. MDE for the purposes of this study is defined as:
        i. The participant indicates at least **five** of the problems listed within the past two weeks (part B on the form), **and**
        ii. **At least one** of the problems must be: symptom 1 (feeling depressed) or symptom 2 (losing interest of pleasure), **and**
        iii. The participant indicates '**yes**' to the question: *Did these problems interfere with your life or activities a lot?*
2. Behavioral activation scale score (BA)
3. Social support (due to the highly skewed nature of this outcome, we dichotomized it into ≥4 versus <4).
4. Mood regulation
5. Decentering
6. Mental health service utilization (i.e., an indicator for whether participants are enrolled in therapy with a counselor and/or currently taking antidepressants). This outcome may be explored in greater detail in a separate paper; however, we plan to examine the difference between arms controlling for baseline mental health service utilization.

Tertiary Outcomes (the following will be explored in a separate paper from the primary results)

1. Pleasant activities
2. Perceived stress
3. Parental cognitions and conduct toward the infant scale
4. Dyadic adjustment scale
5. Flourishing scale
6. Prenatal health behaviors

The outcomes above are tied to validated scales with specific scoring requirements. The scoring instructions will be followed and analyses will proceed on the final, scored variables. In addition, we plan to summarize '**dosage**' of intervention in the active intervention arms as the number of sessions attended (out of six possible). We will also summarize the number of participants that were 'compliant', which is defined as attending **at least four of the six sessions**.

Demographics and Baseline Assessments

The following are specific demographic/baseline assessments of interest for analyses. Primary analyses will adjust for these covariates as we anticipate they will influence outcome. We plan to report both adjusted and unadjusted intervention effect estimates:

1. Race/ethnicity (at the participant level): minority vs. not belonging to a minority category.
2. Whether participant is a first-time mother (i.e., current number of children = 0 at baseline).
3. Currently at risk for depressive episode (i.e., those NOT meeting criteria for MDE at baseline).

4. Primary language of intervention receipt (for MH and HV arms) or primary language in which the participant completed assessments (for control arm).
5. Education: at least some college vs. less than college.
6. Mental health service utilization: we will control for two variables: an indicator for enrollment in counseling with a therapist base baseline, and an indicator for medication for depression at baseline.

Additional demographics that will be summarized in general include:

7. Age.
8. Marital status.
9. Employment status.

Note that some additional exploratory analyses may examine these aforementioned variables as covariates and/or effect modifiers as well. Since these analyses are deemed exploratory, they are not pre-specified. They will be indicated as post hoc in any dissemination materials.

Data Storage

Data will be collected and managed using Research Electronic Data Capture (REDCap) housed at Northwestern University's Clinical and Translational Sciences Institute (CTSA), NUCATS [1]. REDCap is a secure, web-based application designed for research studies that provides an intuitive interface for validated data entry, audit trails for tracking data manipulation and export procedures, and automated export procedures for seamless data downloads to common statistical packages, and procedures for importing data from external sources.

Randomization Methods

The study employs a modified covariate-constrained randomization design [2] at the site level, using unequal (1:3:3; control [C]: mental health [MH] professional delivery of MB: paraprofessional [HV] delivery of MB) allocation, with intention to achieve relative balance in a set of pre-specified potential covariates of interest. There are three variables for which we chose to control imbalance at the site level at baseline through this approach:

1. Percent non-White clients as reported by the site (treated as a continuous variable)
2. Site yearly client volume (also reported by site and treated as a continuous variable)
3. Population density of the site area (continuous variable)

The covariate-constrained method of randomization allows for efficient control over imbalance of multiple covariates at once and is recommended over other methods (i.e., simple randomization or matching) for cluster-randomized trials [2]. The general procedure involves:

1. Enumerating a large subset of possible allocation schemes.
2. Evaluating imbalance for each variable of interest (in this case we have three) for each possible allocation.
3. If the imbalance is acceptable according to some pre-specified criterion, then we save this scheme in a smaller subset of potential allocations for implementation
4. Of those that meet acceptable levels of imbalance, we randomly select one allocation for use in the current study.

We chose the p-value corresponding to the Kruskal-Wallis test (nonparametric version of an Analysis of Variance [ANOVA]) as our criterion for 'balance' in step #3 above. If the p-value for each of the three variables is larger than 0.30 for a given simulated allocation scheme, that particular allocation is deemed 'acceptable'. This criterion is adapted from the 'Minimal Sufficient Balance' principle from Zhao et al. [3] in the individual sequential randomization literature.

In the current study, randomization occurred in three waves for logistical purposes. The first included 14 sites (2 C:6 HV:6 MH); the second included 19 sites (4 C:7 MH:8 HV]; allocation ratio was slightly off in this wave to account for dropout sites); and the third wave included 12 sites (1 C:6 MH:5 HV). Thus, we randomized a total of 45 sites in three waves; randomization of 42 sites was the study goal, reflected in our approved research plan. However, to account for site dropping out in the meantime, we adapted the algorithm to increase the number of sites randomized and also to allow for slightly discrepant allocation ratios than the planned 1:3:3 within each wave.

After randomization, eight programs dropped out, thereby yielding a total of 37 study sites contributing data for analyses. Data collected from all study participants will be used in the analysis. Among the 37 active study sites, 16 HV programs received the MB intervention delivered by mental health professionals, 15 received MB intervention delivered by HV paraprofessionals, and six programs serve as control sites. As an aside, 29 intervention sites ended up implementing at least one cohort (15 MH, 14 HV).

Statistical Methods

Descriptive statistics will summarize baseline characteristics (both site-level and participant-level) overall and by arm. As appropriate, mean ± standard deviation (or median [inner quartile range] will be used in cases of skewed or non-normal empirical distributions) and frequency (proportions) will summarize continuous and categorical data, respectively. Analyses will employ normal theory methodology as appropriate, and in cases of violations of assumptions, transformations and/or nonparametric analyses may be utilized.

Analyses will proceed at the participant level. Primary analyses for QIDS-SR16 score at 24 weeks postpartum will involve all the participants in the study contributing data at this time point. Primary analyses will involve a linear mixed model for continuous outcome with independent fixed effect variables of baseline QIDS-SR16, study arm (three-level factor), and the baseline covariates listed above. We will report both covariate-adjusted and unadjusted estimates for intervention effect. Secondary analysis will be performed on those participants considered at risk at baseline (i.e., excluding the participants who do not meet the criteria for MDE as defined above). We plan to account for clustering effects via inclusion of a random site effect, which will allow for distinction of between and within-site variance. We will also explore inclusion of a random 'facilitator' effect as it is possible participants exposed to the same facilitator may have correlated responses. Intra-cluster correlation coefficients will be estimated via variance components estimates.

Intervention effect will first be evaluated via the adjusted Wald type III test for significant study arm effect at the 5% level of significance. If arm is significant at the 5% level, analyses evaluating the **superiority (Aim #1)** of HV-led intervention vs. control will proceed with Tukey's correction for multiple pairwise hypothesis tests. Assuming the Tukey-adjusted p-value for this comparison falls below the 5% level of significance in favor of the HV-led arm, **non-inferiority (Aim #2)** will further be assessed (using Tukey's correction) via pairwise comparison of adjusted 24-month QIDS in HV-led vs. MH-led arms. Figure1 below depicts the margin and zone of non-inferiority for this comparison.

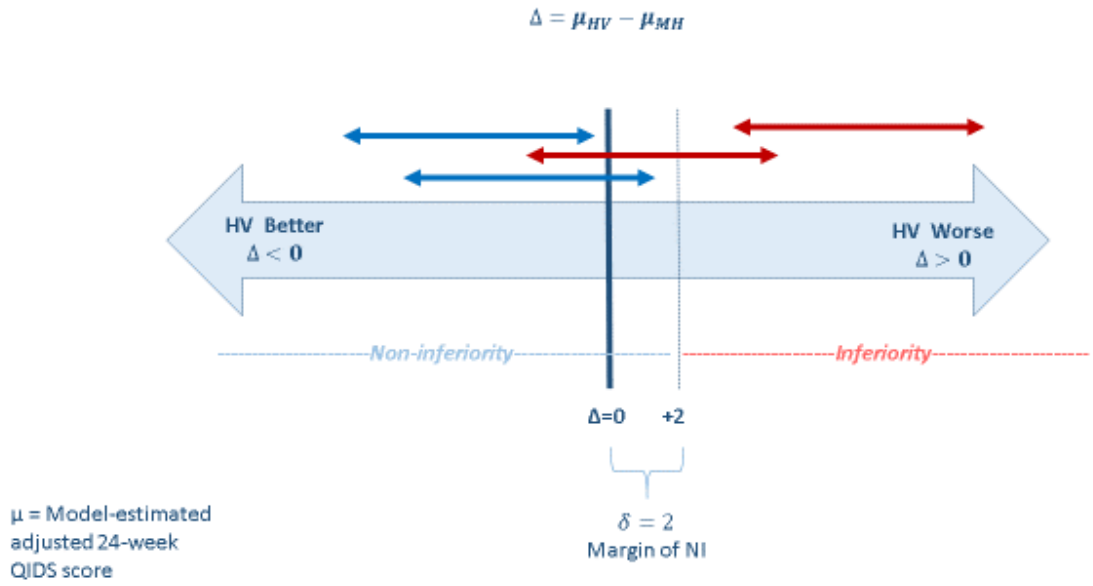

Figure 1. Margin of Non-inferiority (NI)
The double-sided arrows represent 95% confidence intervals for the model-estimated adjusted mean 24-week difference in QIDS-SR16 score. Note that higher score signifies more depressive symptoms; thus if the estimated difference (HV minus MH score) is larger than zero, the HV arm has on average worse depressive symptoms. Since the pre-specified margin of non-inferiority is two units on the QIDS-SR16 scale, we will claim non-inferiority if the upper limit of the adjusted 95% confidence interval for the HV-led arm minus the MH-led arm comparison remains below two. The red arrows indicate scenarios in which we cannot claim non-inferiority and the blue arrows indicate scenarios in which the criterion for non-inferiority is met.

Using the same analytic approach, we will explore the addition of covariate-by-arm interaction terms in the aforementioned model or a three-way covariate-by-arm-by-time interaction term in the longitudinal modelling in order to explore whether effectiveness of intervention varies by patient characteristics (Aim #3). The covariates to be explored include: race/ethnicity, first-time mother, language, education, and use of mental health services at baseline. These analyses are more exploratory in nature, and thus power/sample size considerations do not focus on interaction effects. As a result, we anticipate evaluating interaction effects at the 10% level of significance without adjustment for multiple hypothesis tests. Qualitative data analysis methods will evaluate feasibility and acceptability of MB in each setting (Aim #4), and the details of these analyses will be specified elsewhere.

Additional analyses surrounding Aims #1-#3 will employ longitudinal methods, utilizing all study data at all time points (i.e., inclusion of a fixed study time point effect). Secondary outcomes will all be analyzed in this fashion with the exception being binary response variables (e.g., depression onset). In these cases, appropriate link and distributional assumptions (logit, binomial, respectively) will be specified. For the participants receiving active intervention (either HV-led or MH-led Mothers and Babies), we will examine a 'dose' variable in relation to outcomes. This dose variable will be defined as the number of sessions (out of six possible) attended for an individual participant.

Analytic Dataset

Analyses will be conducted on the modified intent-to-treat (mITT) dataset whereby all participants randomized with data at the 24-week time point will be analyzed according to arm to which they were allocated. We will further perform a sensitivity analysis on the 'as treated' dataset. Those in either active intervention arm will be considered 'treated' if they attend at least four out of the six sessions required for the Mothers and Babies course. For participants with missing data at baseline and non-missing data at follow-up, we plan to impute mean baseline values from the participants' follow-up in order to allow for inclusion in analyses.

Power and sample size considerations allowed for some missing data and, as specified in our approved research plan, we have allowed for up to two clusters to drop out of each arm; however, in the event of large amounts of missing data (i.e., more than 15%), multiple imputation analyses will be explored. We will examine rates of missing data for all variables and determine whether the rates vary by participant characteristics, HV program location, or intervention arm. These summarizations will inform potential biases resulting from missing data. The mixed effects models planned for analysis are generally robust for unbalanced data across study time points. We plan to apply multiple imputation in order to account for missing data; we will impute at least five datasets to generate an estimated average intervention effect. These analyses will again serve as sensitivity analyses to the previously outlined analyses. For participants missing individual items on a particular assessment, we plan to randomly sample from the non-missing empirical distribution for that item. We anticipate less than 5% of participants that would require this imputation of individual items. Again, in the event of large amounts of missing individual items, we will explore multiple imputation methods.

Power and Sample Size Considerations

We initially planned to conservatively enroll 42 sites (six controls, 18 HV-led, 18 MH-led) with a target enrollment of 40 participants per site (1680 total participants). As in any cluster-randomized trial, power and sample size considerations depend heavily on intra-cluster correlation coefficient (ICC) estimates. In general, it is recommended that these calculations account for anticipated ICC as failure to do so in the design phase and lead to an increase in type II error (i.e., result in underpowered studies) [4, 5]. Without previous knowledge of ICC(s) in this population with respect to our primary outcome, we explored a range of ICCs (0.001 to 0.05). Power calculations assume a standard deviation in primary outcome of approximately six points (on the QIDS-SR16), with a meaningful difference corresponding to five points on average across arms. Thus, for the superiority aim, we calculated power based on the average ability to detect at least a five-point mean difference between control sites and HV-led sites. Power calculations assumed a 5%/3 = 1.7% level of significance in order to account for three pairwise comparisons. This is an approach that we deem conservative as this mirrors the Bonferroni correction for multiple hypothesis tests.

We have 38 active sites: six control, 16 paraprofessional home visitor-led sites, and 16 MHP-led sites. We plan to recruit an average of 27 participants per site, allowing for up to 15% attrition (i.e., 23 participants on average for analyses). These assumptions allow for more than 95% power to detect a mean difference of five points in QIDS-SR16 across arms for an ICC of 0.01. Even if one of the six sites drops out, and if ICC is as large as 0.05 (which we deem unlikely), we still anticipate over 85% power for analyses addressing the efficacy aim (Aim #1).

Power for the NI aim will require ability to detect a smaller (margin of NI of two points) mean difference across arms. We anticipate over 90% power to detect a margin of NI of two points on the QIDS-SR16 scale if we assume an ICC of 0.01 and 16 sites in each of the intervention arms with 23 participants for analyses, on average, per site. This allows for up to 15% attrition overall. We anticipate some sites to be over/underperforming and thus unequal representation per site is inevitable. Our hope is that the precautions taken with respect to the randomization algorithm that

attempts to control imbalance in yearly volume and population density will offset biases created by over/underrepresentation for participants at specific sites. While we do not anticipate ICC to be larger than 0.01 [4], we present the required sample size per site in Table 1 below in order to ensure 90% power under the same assumptions for all scenarios explored. We also present sample size requirements in the event that one of the active sites drops out in each of the intervention arms (i.e., 15 sites per arm). Notice that if ICC is larger than 0.02 (which we are not anticipating, although it remains possible), our projected sample size does not allow for 90% power in this case. Since these assumptions are all rather conservative, we argue adequate power for detection of both superiority and non-inferiority. Power calculations for assessment of heterogeneity of intervention effects depending on participant characteristics (Aim #3) require even more assumptions for which we have little information. Thus, we do not necessarily anticipate power to detect specific effects within subgroups or power to detect interaction effects, but we plan to use the analyses outlined here to explore these effects.

Table 1. Required Sample Size for 90% Power in Non-inferiority Aim

ICC	N HV sites	N MH Sites	N per site for analyses	N recruited per site (allowing for 15% attrition)
0.001	16	16	17	20
0.01			20	23
0.02			25	29
0.03			33	39
0.04			49	58
0.05			97	114
0.001	15	15	19	22
0.01			22	26
0.02			28	33
0.03			38	45
0.04			62	73
0.05			168	198

Technical Details

The SAP is subject to version control and we anticipate modifications to analytic plans be documented herein. As in any study, the analytic plan may change due to assumption violations, logistical issues, and/or unexpected empirical distributions of study outcomes. In these cases, the statistical analysis plan will be updated accordingly. All analyses will be performed via SAS version 9.4 or higher (The SAS Institute; Cary, NC) or R version 3.4.1 or higher (The R Foundation for Statistical Computing platform). Example t table shells are included below. Note that these are not binding and may be adapted as needed as formatting and style may be dictated by mode of dissemination or specific target journal(s) for results dissemination.

Timeline for Analyses

To preserve the integrity of the study, no formal analyses will occur until the REDCap database has been locked and all queries/discrepancies resolved; the date of database lock will be documented. Interim descriptive statistics may be generated on an ad hoc basis or for reporting purposes, but no data will be summarized by arm until final analyses unless for quality control purposes. There are no sequential stopping boundaries and interim 'analyses' will not involve formal hypothesis tests. Interim analyses will primarily consist of descriptive statistics (e.g.,

participant status summaries, data completeness, process measures, racial/ethnic breakdown, losses to follow-up, etc.) for study conduct monitoring and as required by PCORI and/or IRB.

Table Shells

Table 1: Baseline Characteristics

	Control	MH	HV	Overall
Age (mean±standard deviation)				
Race / ethnicity				
Employment Status				
Education				
Income				
Primary language				
Born outside of US				
Years in US (Non-natives)				
State of residence				
Planned pregnancy				
Weeks' gestation				
Number of current children				
Currently in counseling w/ therapist				
Meets criteria for MDE				

Table 2: Primary and Secondary Outcome Results

	Control	MH	HV	Overall Result	HV vs. C (Estimate and 95% CI)	HV vs. MH (Estimate and 95% CI)
QIDS Score						
Baseline						
Post-intervention						
Week 12						
Week 24						
Incidence MDE						
Baseline						
Post-intervention						
Week 12						
Week 24						
BA Score						
Baseline						
Post-intervention						
Week 12						
Week 24						
Social Support						
Baseline						
Post-intervention						
Week 12						
Week 24						
Decentering						
Baseline						
Post-intervention						
Week 12						
Week 24						
Mood Regulation						
Baseline						
Week 12						
Week 24						

References

1. Harris, P.A., et al., *Research electronic data capture (REDCap): a metadata-driven methodology and workflow process for providing translational reserach informatics support.* Journal of Biomedical Informatics, 2009. **42**(4): p. 377-381.
2. Ivers, N.M., et al., *Allocation techniques for balance at baseline in cluster randomized trials: a methodological review.* Trials, 2012. **13**(1): p. 120.
3. Zhao, W., M.D. Hill, and Y. Palesch, *Minimal sufficient balance—a new strategy to balance baseline covariates and preserve randomness of treatment allocation.* Statistical methods in medical research, 2015. **24**(6): p. 989-1002.
4. Raab, G.M. and I. Butcher, *Balance in cluster randomized trials.* Statistics in medicine, 2001. **20**(3): p. 351-365.
5. Turner, E.L., et al., *Review of Recent Methodological Developments in Group-Randomized Trials: Part 1—Design.* American journal of public health, 2017(0): p. e1-e9.