ANALYSIS OUTLINE FOR HAPI –[PILOT EFFECTIVENESS]

Outcome variables

Primary endpoint

The primary outcome is mild-to-moderate hypovolemic maternal dehydration during labour, defined dichotomously as a urine specific gravity (USG) measurement greater than 1.015. USG is assessed using a validated digital handheld refractometer (ATAGO PAL-10S, Cat. No. 4410) at the time of labour. This binary definition follows the study's published protocol and reflects a clinically meaningful threshold for dehydration in this population. Although USG is also measured at enrolment, at a 4-week antenatal follow-up visit, and postpartum (within 72 hours, extended to 10 days if needed), the primary endpoint is fixed to the labour timepoint only.

Secondary endpoints

Secondary outcomes are:

1. Heat strain (physiological) – skin temperature measured by non-contact infrared thermometer at the same four timepoints; analysed as a continuous variable (mean values compared between cohorts).
2. Maternal psychological wellbeing – three continuous scores: Women's Wellbeing and Capability Index (WWCI), EQ-5D-5L (including EQ-VAS), and Perceived Stress Scale (PSS-10).
3. Maternal composite clinical outcome – binary indicator of one or more obstetric complications (emergency Caesarean, pre-eclampsia/eclampsia, antepartum/postpartum haemorrhage, maternal sepsis).
4. Neonatal composite clinical outcome – binary indicator of one or more adverse events (stillbirth, preterm birth, low birthweight, neonatal death, or neonatal sepsis).

Primary analysis: propensity score matched comparison

For the primary endpoint (binary dehydration during labour), the primary inference is based on the matched sample generated by the PSM procedure described below. In the matched sample, the average treatment effect on the treated (ATT) is estimated using conditional logistic regression stratified on matched pairs. Results are reported as an odds ratio with 95% confidence interval and two-sided p-value. The same matched pairs are used for all secondary endpoints that lend themselves to pair-level analysis.

For **continuous secondary endpoints** (skin temperature, WWCI, EQ-5D-5L, PSS-10, household temperature measures), we use a **linear mixed model** with matched pair as a random effect. For **binary secondary endpoints** (maternal composite, neonatal composite), we use the same conditional logistic regression as for the primary outcome. Implementation outcomes are analysed using descriptive statistics and thematic

analysis as per the process evaluation plan; they are not subjected to hypothesis testing between cohorts.

Secondary analysis: unmatched comparison

To assess the sensitivity of our results to the matching procedure, we present a secondary, unmatched analysis using the raw pre-intervention and post-intervention cohorts without any adjustment for confounding. For the **primary binary endpoint** (USG > 1.015 during labour), the unmatched analysis uses a **chi-square test**. For **continuous secondary endpoints**, we use an independent-samples **Mann–Whitney U test** (given that these measures are typically non-normally distributed in this setting) and report median differences with Hodges-Lehmann confidence intervals. For binary secondary endpoints, we again use a chi-square test. The unmatched analysis is explicitly labelled as secondary and is not used for primary conclusions. Any divergence between matched and unmatched findings is interpreted as evidence of confounding that the matching procedure successfully addressed.

The Propensity Score Matching (PSM)

Given the quasi-experimental pre-post design, was used to construct comparable groups using propensity score matching (PSM) to reduce selection bias arising from non-random assignment. The pre-intervention cohort serves as the control group ($Z = 0$) and the post-intervention cohort as the treatment group ($Z = 1$). We will estimate the average treatment effect on the treated (ATT) after matching.

A directed acyclic graph (DAG) was used to identify the sufficient set of confounders for matching participants between the pre-intervention (control) and post-intervention (treatment) cohorts. The matching model included age, marital status, site, housing type, education, and employment status, as illustrated in the DAG presented in Figure 1 below.

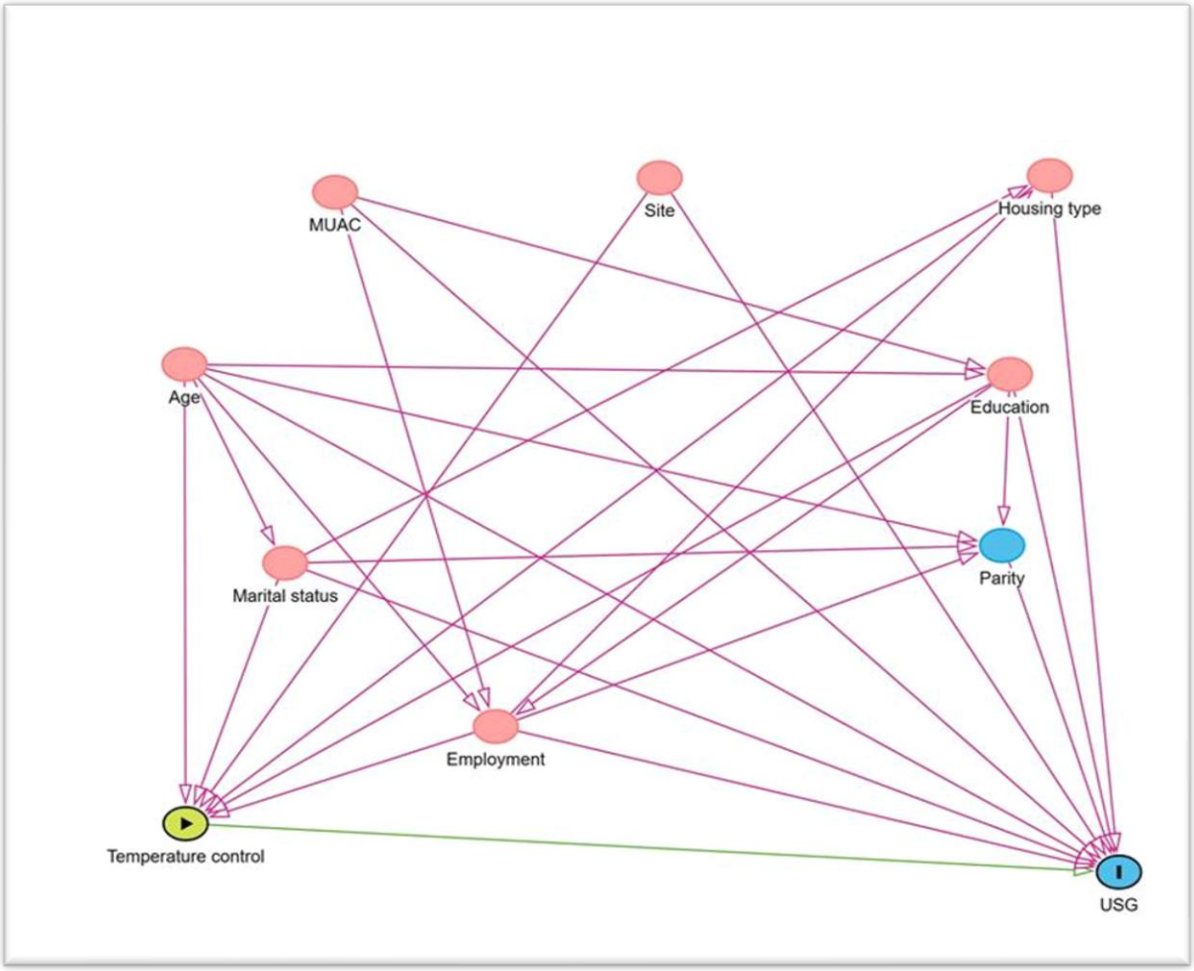

Figure 1: Directed Acyclic Graph (DAG)

Notation and framework

Let $\{(Y_i, Z_i, \mathbf{X}_i)\}_{i=1}^{n}$ denote observations where $Z_i \in \{0,1\}$ indicates cohort assignment, $\mathbf{X}_i \in \mathbb{R}^p$ is a vector of baseline covariates, and Y_i the outcome. Under the potential outcomes framework, we target:

$$\tau_{\text{ATT}} = \mathbb{E}[Y(1) - Y(0) \mid Z = 1].$$

The propensity score $e(\mathbf{X}) = P(Z = 1 \mid \mathbf{X})$ was estimated via logistic regression:

$$\hat{e}(\mathbf{X}_i) = \text{logit}^{-1}(\mathbf{X}_i^{\top}\widehat{\boldsymbol{\beta}})$$

where $\widehat{\boldsymbol{\beta}}$ maximizes the log-likelihood. Covariates included maternal age (continuous), education, employment, marital status, study site, and house type (categorical). Model discrimination was assessed using AUC-ROC, calibration via the Hosmer-Lemeshow statistic, and coefficient stability using variance inflation factors. Propensity scores were logit-transformed as $\ell_i = \log[\hat{e}_i/(1 - \hat{e}_i)]$ to linearize the scale and improve distance-metric properties.

Rather than adopting a fixed caliper (e.g., 0.2 σ), we formalized caliper selection as a bi-objective optimization problem balancing imbalance and retention. Define:

- Imbalance function: $B(c) = \max_{k=1,\ldots,p} | \mathrm{SMD}_k(c) |$, where $\mathrm{SMD}_k(c) = (\bar{X}_{k,1} - \bar{X}_{k,0})/s_{\mathrm{pool},k}$;
- Retention function: $R(c) = m(c)/\min(n_1, n_0)$, where $m(c)$ is the number of matched pairs.

We identify the Pareto frontier $\mathcal{P} = \{c : \nexists\, c' \text{ with } B(c') \leq B(c), R(c') \geq R(c), \text{strict in one}\}$ over $c \in [0.2\sigma_\ell, 1.2\sigma_\ell]$, where $\sigma_\ell = \mathrm{SD}(\ell)$. The optimal caliper c^* is selected as the largest $c \in \mathcal{P}$ satisfying $B(c) \leq 0.1$ (balance-constrained criterion), thereby maximizing sample retention while maintaining acceptable balance.

We implemented nearest-neighbour matching without replacement on the logit propensity score. For each treated unit i, the matched control is:

$$j^*(i) = \arg\min \mathrm{j}: \mathrm{Zj} = 0 \mid \ell_i - \ell_j \mid \text{s.t.} \mid \ell_i - \ell_j \mid \leq c^*.$$

Common support was enforced by restricting matches to the overlapping region of the propensity score distributions. Cross-site matching was permitted to maximise match rates.

Post-matching balance was evaluated using standardized mean differences (target $|\mathrm{SMD}| < 0.1$) and variance ratios. Love plots will visualise balance before and after matching.

For the matched sample, the ATT for binary outcomes (e.g., dehydration status, composite clinical events) will be estimated using ***conditional logistic regression or generalised linear models with robust standard errors***. For continuous outcomes (e.g., USG as a continuous measure, knowledge scores, wellbeing indices), we will use ***linear mixed models with matched pair as a random effect***. All analyses will follow the intention-to-treat principle.

Robustness to unmeasured confounding will be assessed via Rosenbaum bounds, determining the critical Γ at which inference would reverse under hidden bias of magnitude:

$$\frac{1}{\Gamma} \leq \frac{P(Z = 1 \mid \mathbf{X}, U)}{P(Z = 0 \mid \mathbf{X}, U)} \leq \Gamma.$$

Analyses will be conducted in Python 3.10 using scikit-learn for propensity score estimation, scipy for optimal matching, and statsmodels for inference.

Ethical Considerations

The study protocol was approved by the Medical Research Council of Zimbabwe (MRCZ A/3111, approved 12 December 2024). All participants provided written informed consent prior to enrolment. Data were anonymised prior to analysis.